8-2018

Generalized spatiotemporal modeling and causal inference for assessing treatment effects for multiple groups for ordinal outcome.

Soutik Ghosal
University of Louisville

GENERALIZED SPATIOTEMPORAL MODELING AND CAUSAL INFERENCE FOR ASSESSING TREATMENT EFFECTS FOR MULTIPLE GROUPS FOR ORDINAL OUTCOME

By

Soutik Ghosal
B.Sc., Presidency College, 2012
M.Sc., Presidency University, 2014

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August, 2018

GENERALIZED SPATIOTEMPORAL MODELING AND CAUSAL INFERENCE FOR ASSESSING TREATMENT EFFECTS FOR MULTIPLE GROUPS FOR ORDINAL OUTCOME

By

Soutik Ghosal
B.Sc., Presidency College, 2012
M.Sc., Presidency University, 2014

A Dissertation Approved on

July 25, 2018

by the following Dissertation Committee:

Maiying Kong, Ph.D., Dissertation Director

Jeremy Gaskins, Ph.D.

K.B. Kulasekera, Ph.D.

Ritendranath Mitra, Ph.D.

Beatrice Ugiliweneza, Ph.D., MSPH

DEDICATION

This dissertation is dedicated to my maternal grandmother

the late Mrs. Smritikana Chakraborty

without whom this journey would not have started at the very first.

ACKNOWLEDGMENTS

I would like to express my great gratitude to my advisor Dr. Maiying Kong for her constant support and guidance. She not only guided me in my research, but also helped me to shape my future with valuable advices. She made me more confident than I used to be at handling complex problems in research. One cannot hope for a better mentor than her and my research would not have been possible without her guidance.

I would also like to thank Drs. Jeremy Gaskins, Ritendranath Mitra, KB Kulasekera, and Beatrice Ugiliweneza for their time to serve in my dissertation committee and their constructive comments for my dissertation. I specially would like to thank Dr. Gaskins for advising my first project (Chapter 2). I also like to express my gratitude to Dr. John Myers and other members of Child and Adolescent Health Research Design and Support (CAHRDS) Unit of the Department of Pediatrics for providing me the opportunity to work on different pediatric research projects and supporting me financially for the last two years. I am sincerely thankful to all the faculty, students, and administrative staff of the Department of Bioinformatics and Biostatistics for making this journey possible.

Finally, I would like to express my deepest gratitude to my parents, Mr. Sunil Ghosal and Mrs. Tandra Ghosal, who always believed in me and encouraged me throughout my life. Last but not the least, I am very grateful to my dearest friend Debamita for her company, patience, support, and love which made my PhD life extremely smoother and more enjoyable.

ABSTRACT

GENERALIZED SPATIOTEMPORAL MODELING AND CAUSAL INFERENCE FOR ASSESSING TREATMENT EFFECTS FOR MULTIPLE GROUPS FOR ORDINAL OUTCOME

Soutik Ghosal

July 25, 2018

This dissertation consists of three projects and can be categorized in two broad research areas: generalized spatiotemporal modeling and causal inference based on observational data. In the first project, I introduce a Bayesian hierarchical mixed effect hurdle model with a nested random effect structure to model the count for primary care providers and understand their spatial and temporal variation. This study further enables us to identify the health professional shortage areas and the possible impacting factors. In the second project, I have unified popular parametric and nonparametric propensity score-based methods to assess the treatment effect of multiple groups for ordinal outcome. I have conducted different simulation scenarios and compared the performance of those methods. In the third project, I have introduced a generalized spatiotemporal model to identify the antibiotic medication overuse in Kentucky. In this project, I used the Medicaid data to understand the spatial and seasonal variation of the antibiotic overuse for children insured by Kentucky Medicaid.

TABLE OF CONTENTS

PAGE

DEDICATION iii

ACKNOWLEDGMENTS iv

ABSTRACT v

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1: INTRODUCTION 1
- 1.1 Hierarchical Mixed Effect Hurdle Model for Time and Spatially Correlated Count Data and Its Application to Identifying Factors Impacting Health Professional Shortages . . . 1
- 1.2 Comparisons of Average Treatment Effects for Multiple Groups when Outcome is Ordinal and Confounding Exists . . . 2
- 1.3 Generalized Spatiotemporal Additive Model Implemented in R and Its Application to Assessing Overuse of Antibiotics Drugs for Upper Respiratory Tract Infections in Kentucky . . . 2

CHAPTER 2: HIERARCHICAL MIXED EFFECT HURDLE MODEL FOR TIME AND SPATIALLY CORRELATED COUNT DATA AND ITS APPLICATION TO IDENTIFYING FACTORS IMPACTING HEALTH PROFESSIONAL SHORTAGES 3
- 2.1 Introduction . . . 4
- 2.2 Hierarchical mixed effect hurdle model . . . 7
  - 2.2.1 Model . . . 7
  - 2.2.2 Bayesian estimation and inference . . . 11
- 2.3 Case study on HPSA data . . . 14
  - 2.3.1 Data . . . 14
  - 2.3.2 Results . . . 15
- 2.4 Simulation study . . . 18
- 2.5 Conclusion and discussion . . . 20
- 2.6 Tables and Figures . . . 22

CHAPTER 3: COMPARISONS OF AVERAGE TREATMENT EFFECTS FOR MULTIPLE GROUPS WHEN OUTCOME IS ORDINAL AND CONFOUNDING EXISTS 25
- 3.1 Introduction . . . 27

3.2 Data structure and cumulative link models 29
3.2.1 Estimand for ATE for ordinal outcome 30
3.2.2 Parametric approaches to estimate ATE 31
3.3 Proposed GPS-based approaches 34
3.3.1 Generalized propensity score (GPS) models 34
3.3.2 Adjusted U-Statistic for ordinal outcome 35
3.3.3 GPS-based regression 40
3.3.4 GPS-based stratification 41
3.3.5 Covariate balances 42
3.4 Simulation studies 42
3.4.1 Simulation scenarios 44
3.4.2 Simulation results 45
3.5 Case Study 47
3.6 Conclusion and discussion 49
3.7 Tables and Figures 50

CHAPTER 4: GENERALIZED SPATIOTEMPORAL ADDITIVE MODEL IMPLEMENTED IN R AND ITS APPLICATION TO ASSESSING OVERUSE OF ANTIBIOTICS DRUGS FOR UPPER RESPIRATORY TRACT INFECTIONS IN KENTUCKY 58
4.1 Introduction 59
4.2 Generalized spatiotemporal additive model 61
4.3 Study of antibiotic overuse based on Kentucky Medicaid data 62
4.3.1 Background and data set 62
4.3.2 Analysis and results 64
4.4 Conclusion and discussion 65
4.5 Tables and Figures 67

REFERENCES 71

APPENDIX 81
Appendix A 81
A.1 Estimate and variance of ϕ 81
A.2 Model checking and validation 84
A.3 Additional simulation results 88
Appendix B 91
A.4 Other simulation results for ordinal outcome 91
A.5 Score functions used in adjusted U-statistic 101
Appendix C 107
A.6 R code and STAN code for Hierarchical Mixed Effect Hurdle Model for Time and Spatially Correlated Count Data and its Bayesian Analysis 107
A.7 R code for Comparisons of Average Treatment Effects for Multiple Groups when Outcome is Ordinal and Confounding Exists 120

A.8 R code for Generalized Spatiotemporal Additive Model Implemented in R and Its Application to Assessing Overuse of Antibiotics Drugs for Upper Respiratory Tract Infections in Kentucky 157

CURRICULUM VITA 161

LIST OF TABLES

TABLE PAGE

2.1 The variables which may impact the outcome variable (i.e. the number of primary care physicians) 22
2.2 Fixed and random effect parameter estimates and their 95 % credible intervals for binary and count models 23
2.3 Simulation results when the underlying model was NB-SP-ST, but the data was fitted with different models 24
3.1 Rejection rate for the overall test for different methods, where the treatment assignment was generated from OLR, and outcome variable was generated from ordinal logit model (OR1), probit model (OR2), and Box-cox model (OR3), respectively 50
3.2 Observed frequency based on toxic metal exposure and the risk of CKD 50
3.3 The summarized demographic variables across different cadmium and arsenic exposure groups 51
3.4 Superiority score estimates 51
4.1 Summarized description of covariates across levels of antibiotic overuse 67
4.2 Estimated association between antibiotics overuse and sex, race, and medical regions 68
4.3 The top 20 health care providers who prescribed antibiotics for patients diagnosed with URI 68
A1.1 Simulation results when the underlying model was NB-ST, but the data was fitted with different models 89
A1.2 Simulation results when the underlying model was POI-SP-ST, but the data was fitted with different models 90
A1.3 Rejection rate for the overall test for different methods where the treatment assignment was generated from multinomial regression model, and outcome variable was generated from ordinal logit model (OR1), probit model (OR2), and Box-cox model (OR3) respectively 91

LIST OF FIGURES

FIGURE PAGE

2.1 Observed number of primary care physicians at county level (on the log-scale) is shown in panel (a), and the observed number of primary care physicians per 3500 persons is shown in panel (b). If the observed number of primary care physicians per 3500 persons is 6 or more, we label it as "5+". 25
2.2 The illustration of the selected knots 25
2.3 Illustration of different location effects 26
3.1 Power curve for all methods when treatment was generated from OLR 51
3.2 Bias plot for superiority scores when outcome was from Boxcox model and treatment was generated from ordinal logit model 52
3.3 MSE plot for superiority scores when outcome was from Boxcox model and treatment was generated from ordinal logit model 53
3.4 Bias plot for superiority scores when outcome was from ordinal logit model and treatment was from ordinal logit model 54
3.5 MSE plot for superiority scores when outcome was from ordinal logit model and treatment was generated from ordinal logit model 55
3.6 Bias plot for superiority scores when outcome was from ordinal probit model and treatment was generated from ordinal logit model 56
3.7 MSE plot for superiority scores when outcome was from probit model and treatment was generated from ordinal logit model 57
3.8 CKD prognostic status based on ACR and $eGFR$ (Image taken from Levey and Coresh (2012)) 57
3.9 Balance of covariates using weights from multinomial and ordinal logistic regression 58
4.1 KY Medicaid MCO regions (Image taken from Marton et al. (2016)) . 69
4.2 Fraction of antibiotic overuse across different zip codes for children diagnosed with URI based on 2014-2016 Kentucky Medicaid data . . 69
4.3 The complex association between antibiotic overuse and (a) age in months; (b) percentage of poverty level; (c) number of pediatricians; and (d) unemployment rate 70
4.4 Illustration of time trend and seasonal variation 70
A1.1 Estimation of ϕ using two-stage approach 82
A1.2 Illustration of different kinds of Posterior predictive plots 87
A1.3 The power curves for testing overall effect for all methods when treatment was generated from multinomial regression 92

A1.4 Bias plot for superiority scores when response was generated from Box-cox model and treatment was generated from multinomial logistic regression 92
A1.5 Bias plot for superiority scores when response was generated from logit model and treatment was generated from multinomial logistic regression 93
A1.6 Bias plot for superiority scores when response was generated from probit model and treatment was generated from multinomial regression . 94
A1.7 Bias plot for superiority scores when response was generated from Box-cox model, treatment was generated from ordinal, and confounding variables are highly correlated 95
A1.8 Bias plot for superiority scores when response was generated from ordinal logit model, treatment was generated from ordinal logit model, and confounding variables are highly correlated 96
A1.9 Bias plot for superiority scores when response was generated from probit model, treatment was generated from ordinal logit model, and confounding variables are highly correlated 97
A1.10 Bias plot for superiority scores when response was generated from Box-cox model, treatment was generated from multinomial logistic regression model, and confounding variables are highly correlated 98
A1.11 Bias plot for superiority scores when response was generated from ordinal logistic regression model, treatment was generated from multinomial logistic regression model, and confounding variables are highly correlated 99
A1.12 Bias plot for superiority scores when response was generated from probit model, treatment was generated from multinomial logistic regression model, and confounding variables are highly correlated 100

CHAPTER 1

INTRODUCTION

1.1 Hierarchical Mixed Effect Hurdle Model for Time and Spatially Correlated Count Data and Its Application to Identifying Factors Impacting Health Professional Shortages

Count data is common in many fields such as public health. Hurdle models have been developed to model count data where zero count could be either inflated or deflated. However, when data is repeatedly collected over time and the data is also spatially correlated, it is very challenging to model the data appropriately. For example, to study health professional shortage areas, the numbers of primary care physicians along with other demographic characteristics are collected at county level in the USA and over different years. Since the data is repeatedly collected over time, counties are nested within the state, and adjacent counties are geographically correlated, the dependence structure of the data is very complex. We develop a Bayesian hurdle model with multi-layered random effects to incorporate this complex structure. We use a time-varying random effect for each state to capture the time-effect at the state level, and a thin plate spline to capture the spatial correlation across different counties. We use Stan to obtain samples from the posterior distributions for inference. By using the proposed model, we are able to identify the important factors which impact the health professional shortages. Simulation studies also confirm the effectiveness of the model.

1.2 Comparisons of Average Treatment Effects for Multiple Groups when Outcome is Ordinal and Confounding Exists

Ordinal data are very common in clinical fields, and it is very important to accurately assess the average treatment effects from two or more treatments. Randomized controlled trial (RCT) is considered as a gold standard to estimate the treatment effect. Many popular parametric and non-parametric approaches are developed to assess the treatment effect for RCT. However, RCT may not be always feasible due to ethics, cost, and patients' preferences. With the availability of the observed data in natural health care setting, estimating the average treatment effect based on the observational studies becomes more practical. In the observational studies, the confounding covariates often exist, and the statistical methods developed for RCT may not be suitable anymore. In this project, we investigate parametric and non-parametric methods to compare treatment effects among multiple groups when outcome is ordinal. We use the superiority score as measure of a treatment effect between two groups. We extend the parametric approaches such as ordinal logistic regression and nonparametric method such as adjusted U-statistic to compare the treatment effect in the presence of confounding covariates. A case study is provided to study the effect of cadmium and arsenic on chronic kidney disease based on the National Health and Nutrition Examination Survey (NHANES) 2011-2014 data set.

1.3 Generalized Spatiotemporal Additive Model Implemented in R and Its Application to Assessing Overuse of Antibiotics Drugs for Upper Respiratory Tract Infections in Kentucky

Antibiotics are special types of antimicrobial drugs which are used in the treatment and prevention of various bacterial infections. Over the year use of antibiotics has

increased in the United States, especially in Kentucky. Kentucky has the second highest antibiotic prescription rate in the entire nation. Recent studies have shown that a significant fraction of these prescriptions is unnecessary. For example, patients suffer from upper respiratory tract infections (URI), which do not need to be treated with antibiotics unless there are some chronic conditions or competing diseases. However, based on Kentucky Medicaid data 2014-2016, more than 50% of the visits with URI diagnosis had antibiotics prescriptions. In this study, we investigate whether the antibiotic prescriptions for URI have varied geographically and whether the antibiotic prescriptions have changed over time with possible seasonal variation. We also investigate whether antibiotic prescription is related to demographics and socio-economic conditions. We construct a generalized additive model (GAM) which can be implemented in R and be used to address the study questions. We present the GAM and the R-code in this study, and illustrate how to use the R-code to facilitate the study of the overuse of the antibiotics prescriptions for URI.

CHAPTER 2

HIERARCHICAL MIXED EFFECT HURDLE MODEL FOR TIME AND SPATIALLY CORRELATED COUNT DATA AND ITS APPLICATION TO IDENTIFYING FACTORS IMPACTING HEALTH PROFESSIONAL SHORTAGES

2.1 Introduction

Count data is common in many fields such as in public health and civil and industrial engineering (Lord et al., 2005; Xu et al., 2014; Zhang et al., 2016). When count data has an unusually large number of zeroes, traditional count models (i.e., log-linear models) based on Poisson and Negative Binomial distributions fail to fit the data appropriately. When the number of observed zeros are beyond what Poisson model or Negative Binomial distribution can describe, we call this zero-inflated count data. On the contrary, when the observed zeros are less than what a Poisson or Negative Binomial distribution can describe, we call this zero deflated count data. Two commonly used models to analyze zero-inflated count data are zero-inflated Poisson (ZIP) model (Hur et al., 2002; Lee et al., 2006; Shankar et al., 1997; Wang et al., 2002) and zero-inflated Negative Binomial (ZINB) model (Lim et al., 2013; Mullahy, 1986; Yau et al., 2003). In the zero-inflated model, a zero could come either from a degenerated distribution at zero or from an ordinary count distribution, such as Poisson or Negative Binomial distribution. ZIP model is one of the most used models for fitting zero-inflated count data, however it becomes less effective if the data has a larger vari-

ance than the expected mean value, which is known as the problem of overdispersion. ZINB model is more effective for zero-inflated count data with overdispersion. These methods are suitable for zero-inflated count data but not for zero-deflated count data (Ridout et al., 2001).

In the literature, the hurdle model has been developed to model data with either zero-inflated counts or zero-deflated counts. The hurdle model uses a two-stage modeling process. The first stage models a binary variable of whether the count falls below or above the hurdle (i.e., zero versus positive count), and the second stage uses a truncated count distribution to model the observations above the hurdle. The zeros under the hurdle can be more or less likely than would be predicted under the (untruncated) count distribution. The popular distributions for the second stage are truncated Poisson or Negative Binomial distributions, although other discrete distributions have been used as well (Angers and Biswas, 2003; Choo-Wosoba et al., 2018; Xie et al., 2014). The hurdle model with truncated Poisson may not be appropriate for count data with overdispersion, while the hurdle model with truncated Negative Binomial distribution includes an overdisperion parameter and provides a more flexible choice.

The zero-inflated and the hurdle models generally assume that the observations are independent. When the observed count data are correlated and have excess zero counts, zero-inflated random effect models and zero-inflated generalized estimating equation (GEE) models are often used. The zero-inflated random effect models use random effects to capture the correlations of the observations from an experimental unit (Hall, 2000; Min and Agresti, 2005), while the zero-inflated GEE models (Dobbie and Welsh, 2001; Hall and Zhang, 2004; Kong et al., 2015) use a correlation structure matrix to capture the possible correlations of observations.

In our study of health professional shortage areas (HPSAs), the number of primary care physicians in each county is considered as the outcome variable, which

is collected yearly from 2007-2012. The county level characteristics which impact the outcome variable include population size, average education level, poverty level, racial makeup of the county, and others and are presented in Table 2.1. The data exhibits a complex dependence structure. Counties are nested within states, and adjacent counties are considered to be geographically correlated (Aktekin and Musal, 2015; Musal and Aktekin, 2013). Additionally, the observations from counties within a state are likely to be impacted by state-level policies, and adjacent counties could also be geographically correlated. Since the data are repeatedly collected over time, the observations for a county over different years are not independent. With such a complex data structure, the currently available zero-inflated random effect models and GEE models can not be directly applied. We do not have a priori knowledge on whether to expect zero counts to inflate or deflate relative to the count model, and we have independent interest in the factors that lead to counties with no medical professionals; hence, a hurdle model is more appropriate than a zero-inflated choice.

We first consider the number of primary care physicians per county in log-scale in Figure 2.1a. Due to the skewness, we plot the number of primary care physicians on a $\log(x+1)$ scale (natural logarithm) where the number of physicians in each county is obtained from the mean of the counts from 2007-2012. Clearly, many areas in the Great Plains and other central regions of the country have no or few physicians per county, while coastal and other metropolitan areas contain counties with a high number of physicians. We also plot the number of primary care physicians per 3500 residents to identify the health professional shortage areas in Figure 2.1b. Again, many of the health professional shortage areas appear in the Great Plains region. 6.86% of the US counties are completely unserved by the medical community (zero counts), and the median count of primary care physicians across the data is 12. The maximum count is 9211 which indicates that the data may be overdispersed. To model data with all these considerations, we develop a hierarchical mixed effect

hurdle model for time and spatially correlated count data. In the model, we use a time-varying random effect for each state to capture the dependence of the observations from the same state across different years, and a thin-plate spline model to capture the spatial correlation across adjacent counties. We assume that the number of primary care physicians could be zero-inflated for some areas with poor social economical environment, and the count could be zero-deflated for the areas with excellent social economical environment.

2.2 Hierarchical mixed effect hurdle model

2.2.1 Model

Let us denote Y_{ijk} as the count (e.g., the number of primary care physicians) and X_{ijk} as the vector of covariates for the j^{th} county in the i^{th} state and k^{th} year, where $i = 1, \ldots, N$; $j = 1, \ldots, n_i$; $k = 1, \ldots, T$. We define p_{ijk} as the probability that the count Y_{ijk} is non-zero count (i.e., positive count):

$$Pr[Y_{ijk} > 0 | X_{ijk}] = p_{ijk}, \tag{2.1}$$

and

$$Pr[Y_{ijk} = 0 | X_{ijk}] = 1 - p_{ijk} \ . \tag{2.2}$$

Let us assume that non-zero counts come from a truncated Negative Binomial distribution, where the (untruncated) Negative Binomial distribution is parameterized to have mean μ_{ijk} and dispersion parameter ϕ with variance $\mu_{ijk} + \phi\mu_{ijk}^2$. The

distribution for Y_{ijk} can be written as

$$P[Y_{ijk} = y|X_{ijk}] = \begin{cases} 1 - p_{ijk}, & \text{if } y = 0; \\ p_{ijk} \dfrac{\frac{\Gamma(y+\frac{1}{\phi})}{y!\Gamma(\frac{1}{\phi})}(\frac{\phi\mu_{ijk}}{1+\phi\mu_{ijk}})^y(\frac{1}{1+\phi\mu_{ijk}})^{\frac{1}{\phi}}}{1-\left(\frac{1}{1+\phi\mu_{ijk}}\right)^{\frac{1}{\phi}}}, & \text{if } y > 0. \end{cases} \tag{2.3}$$

We use a logistic regression model for p_{ijk}, and a log-linear model for the mean μ_{ijk} of the Negative Binomial distribution. The logistic regression model for the binary outcome of whether $Y_{ijk} > 0$ is specified as

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = X'_{ijk}\boldsymbol{\beta} + a(s_{1ik} + s_{2ij}), \tag{2.4}$$

and the log-linear model for counts is written as:

$$\log(\mu_{ijk}) = X'_{ijk}\boldsymbol{\gamma} + s_{1ik} + s_{2ij} \ . \tag{2.5}$$

Here, β and γ are fixed effects for the binary model and the count model, respectively. Note that $T-1$ dummy variables for T different years are also included as part of the fixed effects to capture nationwide time effects. s_{1ik} $(i = 1, \ldots, N;\ k = 1, \ldots, T)$ is the random effect for the i^{th} state at k^{th} year to capture the heterogeneity due to state and time variation, and s_{2ij} $(i = 1, \ldots, N; j = 1, \ldots, n_i)$ is the random effect for the j^{th} county in the i^{th} state to capture the county-level (spatial) heterogeneity. For a fixed year k, s_{1ik} describes the variation across states within k^{th} year beyond that explained by the predictors in the fixed effect components. For instance, state-wide laws and policies that impact the number of doctors may be captured by these terms. The vector $\boldsymbol{s_{1i\cdot}} = (s_{1i1}, \ldots, s_{1iT})^T$ captures the time effects across T years for i^{th} state, and an unstructured correlation for $\boldsymbol{s_{1i\cdot}}$ is assumed. By averaging the random effects over time $\frac{1}{T}\sum_{k=1}^{T} s_{1ik}$, we can get an interpretable overall effect of i^{th} state

(beyond what is captured in the fixed effects) to compare across states.

We have shared the total random effect contribution $(s_{1ik} + s_{2ij})$ in the binary model and count model because we expect that states/counties/year that are more likely to produce a non-zero count are also more likely to produce a higher count. We used the multiplying factor a in the binary model to rescale the random effects since the magnitude on the logit link in equation (2.4) may differ from the log-link function in equation (2.5).

We assume that the random effects for the i^{th} state across the T time points follow a normal distribution with mean $\mathbf{0}$ and variance Ψ. That is,

$$\boldsymbol{s_{1i\cdot}} \sim N_T(\mathbf{0}, \Psi), \;\; i = 1, \ldots, N. \tag{2.6}$$

Here Ψ is a $T \times T$ unstructured covariance matrix, in which the diagonal entries capture the state-level variation across different time points, and the off-diagonal entries capture the covariance of two observations from the same state but different years.

The random effect s_{2ij} $(i = 1, \ldots, N; j = 1, \ldots, n_i)$ describes the county level heterogeneity. We expect that the observations from adjacent counties are more correlated than the observations from counties far from each other. So we use thin-plate splines to capture the county-level heterogeneity and to model the spatially correlated data. To use the thin-plate spline technique, we first chose L knots, $\tau_1, \tau_2, \ldots, \tau_L$, where $\tau_l = (\tau_{l_1}, \tau_{l_2})$ represents the latitude and longitude of the l^{th} knot τ_l. Initially those knots were selected by choosing combinations of 15 equidistant latitudes from north to south, and 20 equidistant longitudes from east to west resulting in $L = 300$ equidistant locations in or close to the border of the United States. We then only keep those locations for which there is at least one county center within 100 miles of the knot locations, and we ended up having 211 knots (see Figure 2.2). For a county whose center is at latitude and longitude $w = (w_1, w_2)$, we use the following Gaussian

radial basis functions to form a basis for the thin-plate splines:

$$g_l(w) = e^{-\frac{1}{B}\delta(w,\tau_l)}, \quad l = 1, \ldots, L. \tag{2.7}$$

Here, $\delta(w, \tau_l)$ is the distance (in miles) between the county with the central location at w and the knot τ_l. B is the bandwidth (Braun and Huang, 2005), which is taken as a fixed value. In the case study in Section 3 and simulation studies in Section 4, we set B=50. Using the basis function in equation (2.7), we can write the county effect from j^{th} county in the i^{th} state as linear combination of these basis functions. That is,

$$s_{2ij} = \sum_{l=1}^{L} g_l(w_{ij})\xi_l = Z_{ij}^{'}\boldsymbol{\xi}$$

where $w_{ij} = (w_{ij,1}, w_{ij,2})$, which is the central location of the j^{th} county in the i^{th} state, and $Z_{ij} = (g_1(w_{ij}), \ldots, g_L(w_{ij}))^{'}$ is the vector of basis function values for the j^{th} county in the i^{th} state. After selecting knot locations and bandwidth, Z_{ij} is calculated from the observed data and fixed throughout the analysis. The spline coefficients $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)^{'}$ are unknown and need to be estimated from the data. $\boldsymbol{\xi}$ is usually assumed to follow multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma_\xi^2 I$. If we denote the spline design matrix $\boldsymbol{Z} = (Z_{ij}^{'})$ $(i = 1, \ldots, N; j = 1, \ldots, n_i)$, then the county level spatial random effect $\boldsymbol{S_2} = (s_{2ij})$ $(i = 1, \ldots, N; j = 1, \ldots, n_i)$ can be expressed as $\boldsymbol{Z\xi}$. Marginalizing over $\boldsymbol{\xi}$, the joint distribution of the spatial effect $\boldsymbol{S_2}$ follows $MVN(\mathbf{0}, \sigma_\xi^2 \boldsymbol{Z}\boldsymbol{Z}^{'})$. An alternative option to model the county-level correlation is to use county-level correlated random effects. However, we have found this alternative approach is numerically unstable with some random effects diverging to negative infinity.

Note that variation among counties from different states are explained by $\boldsymbol{S_2}$,

whereas variation across state and time are explained by $\boldsymbol{S_1} = (\boldsymbol{s_{1i\cdot}})$ $(i = 1, \ldots, N)$. Hence, the overall random effect term $\boldsymbol{S_1} + \boldsymbol{S_2}$ determines the dependence pattern across years, across states, and across their counties. The covariance/correlation between particular combinations of state/county/year can be determined from the relevant elements of Ψ and $\sigma_\xi^2 \boldsymbol{Z}\boldsymbol{Z}'$.

2.2.2 Bayesian estimation and inference

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \Psi, \sigma_\xi^2, a, \phi)$ denote the parameters to be estimated. It is very challenging to obtain estimates for all the parameters in $\boldsymbol{\theta}$ under a frequentist estimation approach due to the difficulty in approximating the marginalized likelihood which is obtained from integrating out the random effects. Instead, we seek Bayesian approach to obtain the estimates for all the parameters including random effects using an iterative algorithm that does not require marginalizing out the random effects. We, therefore, assign a prior distribution for each parameter. We use relatively non-informative priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$: $\boldsymbol{\gamma} \sim MVN(\mathbf{0}, 100 I_{m\times m})$ and $\boldsymbol{\beta} \sim MVN(\mathbf{0}, 100 I_{m\times m})$, where m is the number of covariates specified in the model. For the random effect scaling coefficient, we use $a \sim N(0, 100)$. For the covariance matrix of the state-time random effect $\boldsymbol{S_1}$, we use $\Psi \sim Inverse\ Wishart(\nu, \Sigma)$, where $\nu = T + 1$ and $\Sigma = \frac{1}{T} I_{T\times T}$ with $T = 6$. σ_ξ, the standard deviation of the spline coefficients, has a half-Cauchy prior with scale of 0.1, which is common choice to induce some shrinkage of the regression coefficients in hierarchical models (Gelman, 2006). For the NB overdispersion parameter, we use a uniform prior on ϕ with wide support: $\phi \sim Unif(0, 100)$.

Note that $\boldsymbol{S_1} = (s_{1ik})$ $(i = 1, \ldots, N; k = 1, \ldots, T)$ denotes the set of state-time effects and $\boldsymbol{S_2} = (s_{2ij})$ $(i = 1, \ldots, N; j = 1, \ldots, n_i)$ the set of spatial/spline effects. The posterior distribution for $\boldsymbol{\theta}, \boldsymbol{S_1}$ and $\boldsymbol{S_2}$ given $\boldsymbol{Y} = (y_{ijk})$ $(i = 1, \ldots, N; j = 1, \ldots, n_i; k = 1, \ldots, T)$ can be written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{S_1}, \boldsymbol{S_2}|\boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{S_1}, \boldsymbol{S_2}) f(\boldsymbol{S_1}|\boldsymbol{\theta}) f(\boldsymbol{S_2}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta})$$
$$= f(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{S_1}, \boldsymbol{S_2}) f(\boldsymbol{S_1}|\Psi) f(\boldsymbol{S_2}|\sigma_\xi^2) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\gamma}) \pi(\sigma_\xi) \pi(\Psi) \pi(a) \pi(\phi). \tag{2.8}$$

Here $f(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{S_1}, \boldsymbol{S_2})$ is the product of data-likelihood $f(y_{ijk}|\boldsymbol{\theta}, s_{1ik}, s_{2ijk})$ given by equation (2.3). $f(\boldsymbol{S_1}|\Psi)$ is given by the product of $f(\boldsymbol{s_{1i\cdot}}|\Psi)$ over index i where $f(\boldsymbol{s_{1i\cdot}}|\Psi) = (2\pi)^{-\frac{T}{2}} |\Psi|^{-\frac{1}{2}} e^{-\frac{1}{2} \boldsymbol{s_{1i\cdot}}' \Psi^{-1} \boldsymbol{s_{1i\cdot}}}$. $f(\boldsymbol{S_2}|\sigma_\xi^2)$ is given by $\boldsymbol{S_2} = \boldsymbol{Z\xi} \sim MVN(\boldsymbol{0}, \sigma_\xi^2 \boldsymbol{Z}\boldsymbol{Z}^T)$. The rest in terms of $\pi(\cdot)$ are defined from their respective priors.

Sampling from the joint posterior distribution is implemented by the statistical software `Rstan` which is the R interface to `Stan` (Carpenter et al., 2016). The `Stan` language is used to carry out full Bayesian statistical inference using a Hamiltonian Monte Carlo (HMC) scheme. `Stan` provides an automated platform for Bayesian inference that only requires the user to input a hierarchical model structure, and `Stan` develops a fast sampling scheme to provide posterior samples from the model.

When the dispersion parameter ϕ is included in the sampling in `Stan` from the posterior distribution in equation (2.8), we found that the MCMC scheme tends to be computationally burdensome, and the posterior commonly gets stuck in local modes. Hence, we adopt a pseudo-empirical Bayes technique to estimate the dispersion parameter ϕ. To that end, we consider the marginal posterior $\pi(\phi|\mathbf{Y})$ for ϕ given the data $\mathbf{Y}$ using a Newton-Raftery style estimator (Newton and Raftery, 1994). Letting $\boldsymbol{\theta}_-$ denote the vector of parameters excluding ϕ, it follows that

$$\begin{aligned}\frac{1}{\pi(\phi|\mathbf{Y})} &= \frac{m(\mathbf{Y})}{f(\mathbf{Y},\phi)} = \frac{m(\mathbf{Y})}{\pi(\phi)}\int \frac{\pi(\boldsymbol{\theta}_-|\phi)}{f(\mathbf{Y}|\phi)}d\boldsymbol{\theta}_- = \frac{m(\mathbf{Y})}{\pi(\phi)}\int \frac{\pi(\boldsymbol{\theta}_-,\mathbf{Y}|\phi)/f(\mathbf{Y}|\phi)}{\pi(\boldsymbol{\theta}_-,\mathbf{Y}|\phi)/\pi(\boldsymbol{\theta}_-|\phi)}d\boldsymbol{\theta}_- \\ &= \frac{m(\mathbf{Y})}{\pi(\phi)}\int \frac{\pi(\boldsymbol{\theta}_-|\mathbf{Y},\phi)}{f(\mathbf{Y}|\boldsymbol{\theta}_-,\phi)}d\boldsymbol{\theta}_- \\ &= \frac{m(\mathbf{Y})}{\pi(\phi)}E\Big[\, f(\mathbf{Y}|\boldsymbol{\theta}_-,\phi)^{-1}\,|\mathbf{Y},\phi\Big], \end{aligned} \tag{2.9}$$

where the final expectation is with respect to the posterior distribution $\pi(\boldsymbol{\theta}_-|\mathbf{Y},\phi)$. As the prior for ϕ is constant and the marginal likelihood of $\mathbf{Y}$ (i.e., $m(\mathbf{Y})$) does not depend on ϕ, we conclude that

$$\pi(\phi|\mathbf{Y}) \propto \Big\{E\Big[\, f(\mathbf{Y}|\boldsymbol{\theta}_-,\phi)^{-1}\,|\mathbf{Y},\phi\Big]\Big\}^{-1}.$$

Our proposed pseudo-empirical Bayes approach can be carried out by the following multistage technique. In the first stage, we use `Stan` to obtain a posterior sample from the model with a fixed dispersion parameter $\phi = \phi_{(0)}$. Typically, we begin with $\phi_{(0)} = 0$, corresponding to the limiting case of the Poisson distribution. To maximize $\pi(\phi|\mathbf{Y})$, we replace the expectation $E\Big[\, f(\mathbf{Y}|\boldsymbol{\theta}_-,\phi)^{-1}\,|\mathbf{Y},\phi\Big]$ with $E\Big[\, f(\mathbf{Y}|\boldsymbol{\theta}_-,\phi)^{-1}\,|\mathbf{Y},\phi_{(0)}\Big]$. This expectation can be approximated using the posterior sample fit with fixed $\phi = \phi_{(0)}$, and we estimate the (un-normalized) marginal posterior as a function of ϕ by

$$\hat{\pi}(\phi|\mathbf{Y}) = \Big\{E\Big[\, f(\mathbf{Y}|\boldsymbol{\theta}_-,\phi)^{-1}\,|\mathbf{Y},\phi_{(0)}\Big]\Big\}^{-1} \simeq \left\{\frac{1}{G}\sum_g e^{-\sum_{i,j,k} l(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})}\right\}^{-1}. \tag{2.10}$$

Here, $l(\boldsymbol{\theta}_-,\phi|y_{ijk})$ represents the log-likelihood for the parameters $\boldsymbol{\theta}_-$ and ϕ for observation y_{ijk}, and $\boldsymbol{\theta}_-^{(g)}$ represents the parameter sample in the g^{th} MCMC iteration, which has distribution $\boldsymbol{\theta}_-^{(g)} \sim \pi(\boldsymbol{\theta}_-|\mathbf{Y},\phi_{(0)})$ $(g = 1,\ldots,G)$. To update ϕ, we must maximize (2.10), and we set $\phi = \phi_{(1)}$ to be this value.

In the next stage, we again obtain an MCMC sample for $\boldsymbol{\theta}_-$ from `Stan` using the hierarchical mixed Negative Binomial hurdle model with the fixed dispersion parameter at $\phi = \phi_{(1)}$. We re-estimate ϕ by maximizing (2.10) using this new posterior sample from $\pi(\boldsymbol{\theta}_-|\mathbf{Y}, \phi_{(1)})$ and calling the maximum $\phi_{(2)}$. In general, we continue in this way multiple times the estimates of ϕ stabilize at a final value $\hat{\phi}$. Our experience indicates only two or three such steps are typically needed. Inference of the remaining parameters in θ is based on the MCMC sample from the final stage of ϕ estimation.

To simplify maximization of (2.10) in the intermediate steps, we often replace the average over G iterations with an evaluation at the posterior mean $\hat{\boldsymbol{\theta}}_-$. This is equivalent to finding the ϕ that maximizes $\hat{l}(\phi|\mathbf{Y}) = \sum_{i,j,k} l(\hat{\boldsymbol{\theta}}_-, \phi|y_{ijk})$. An illustration of the estimation of ϕ is shown in Appendix A.1.

To represent the precision of the ϕ estimates, we constructed a Wald type 95% credible interval of form $(\hat{\phi} - 1.96 * \sqrt{\hat{s}^2_\phi}, \hat{\phi} + 1.96 * \sqrt{\hat{s}^2_\phi})$, where $\hat{s}^2_\phi$ is an estimate of the variance of ϕ. The variance estimator is obtained by taking the negative inverse of the 2^{nd} derivative of the estimate of $\log \pi(\phi|\mathbf{Y})$ (equation (2.10)) and evaluating it at $\phi = \hat{\phi}$. The numerical form of the variance is given in the Appendix A.1.

To improve the computational efficiency, it is important to choose the initial values carefully. The initial values at the first (Poisson) stage for the fixed effect coefficients β and γ are obtained respectively from the MLEs of a logistic model (for the zeroes) and log-linear model (for the counts) ignoring the random effects. Since we use multistage approach to estimate ϕ, we initialize the parameters in $\boldsymbol{\theta}_-$ in the later stages by using the posterior sample from the previous `Stan` fit. The 95% credible interval for each parameter in $\boldsymbol{\theta}_-$ is constructed as 2.5% and 97.5% percentile of the posterior samples.

2.3 Case study on HPSA data

2.3.1 Data

It is generally known that access to health care is a strong predictor of physical health (Andersen, 1995). In order to improve health care service, the Health Resources and Services Administration (HRSA) (`https://www.hrsa.gov/`) established a criterion for identifying health professional shortage areas (HPSAs). HPSA categorization is calculated by the ratio of the number of primary care physicians to the population size. An area is considered to be a primary care HPSA if the ratio is 1:3500 or less, indicating an insufficient capacity of existing primary care providers.

In this section, we apply our proposed model to identify the factors which may impact the number of primary care physicians in a particular area. Although we do not directly model whether an area is categorized as an HPSA, the factors in our model associated with higher chances of zero primary care physicians and/or lower counts of primary care physicians are likely to be the same factors associated with HPSA shortages. In this section, we apply our proposed model to identify the factors that impact the number of primary care physicians in a county. The data set has information on the 48 contiguous states and the District of Columbia. There are 3105 counties or county equivalents in the data set with observations from years 2007 to 2012 resulting in 18,630 observations. The response variable is the number of physicians at each county at a specified year. The data set also includes covariates such as location characteristics (e.g., rural versus urban or suburban), population demographics (e.g., age, sex, race, and education) and local economic conditions (e.g., labor force participation, unemployment rate, health insurance coverage, poverty level, and per-capita income). The complete list of covariates and their descriptions are found in Table 2.1.

2.3.2 Results

We used `Rstan` to analyze the HPSA data by using the proposed model and estimation procedure. To assist interpretation, all covariates are standardized to mean 0 and variance 1. In `Rstan`, we ran 3 parallel chains with 1000 iterations for each chain. We discarded the first 500 samples as burn-in in each chain resulting in 1500 total samples for each parameter. The dispersion parameter ϕ stabilized in 3 steps and the corresponding estimates are respectively 0.086, 0.087 and 0.087. To examine the convergence within the final MCMC sample, we monitored the potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992), as well as the trace plots for the fixed effect coefficients. The initial stage of fitting the Poisson model is computationally slow due to inconsistency between Poisson model and the over-dispersed data. However, once ϕ is fixed at a convergent value, the computation takes about one-fourth of the time of the initial Poisson stage.

The estimates of the fixed parameters as well as their equal tail 95% posterior intervals are presented in Table 2.2. Estimates whose 95% credible intervals (CI) exclude zero are highlighted in bold, and the predictors whose CIs for both the binary and count models don't cover zero are also highlighted. From Table 2.2, the population size of the county is strongly associated with the number of primary care physicians. A one unit increase in the log-population corresponds to an adjusted log-odds ratio of 3.91 for having at least one physician and a five-fold (i.e., $e^{1.687} \approx 5$) increase in the number of primary care physicians for those counties with at least one physician. Highly educated counties (large values of 18+ years education) and counties with a medical school show strong positive associations with the presence of and number of primary care providers. Other predictors that are positively associated with having more primary care physicians include the following: percentage of African Americans, percentage of Hispanics, percentage of females, urban influence (i.e. how urban the county is), per capita income, percentage of 65+ older popula-

tion, high birth rates, and high death rates. Interestingly, the percentage of Asian population has a negative correlation to the outcome variable, although this may be related to high correlation between predictors. Some variables were significant in only one of the models. For instance, elevated state income tax is associated with increase of the number of non-zero counts, and elevated labor percentage with an increase in the number of primary care physicians in the county. The estimates of the variance components and their 95% credible intervals are presented in Table 2.2. The estimate of the scale parameter a indicates that the random component has a larger impact in the binary model than the count model. The estimate of Ψ which captures the variation across states and years is given by

$$\begin{bmatrix} 0.035 & 0.029 & 0.028 & 0.027 & 0.026 & 0.027 \\ \mathbf{0.833} & 0.034 & 0.028 & 0.027 & 0.025 & 0.027 \\ \mathbf{0.826} & \mathbf{0.831} & 0.033 & 0.027 & 0.025 & 0.026 \\ \mathbf{0.813} & \mathbf{0.819} & \mathbf{0.827} & 0.032 & 0.025 & 0.026 \\ \mathbf{0.803} & \mathbf{0.808} & \mathbf{0.816} & \mathbf{0.814} & 0.029 & 0.025 \\ \mathbf{0.806} & \mathbf{0.812} & \mathbf{0.812} & \mathbf{0.819} & \mathbf{0.812} & 0.031 \end{bmatrix}.$$

Here the entries in the upper triangle of the matrix are the estimated variance-covariance matrix, and the elements in the lower triangle (in bold) are the corresponding correlations. The high correlations in Ψ for the state-level random effects $\boldsymbol{s_{1i\cdot}}$ indicate that the states with a higher/lower than predicted number of primary care physicians tend to stay higher/lower throughout the study.

We present the summarized results in Figure 2.3 . Figure 2.3a displays the estimated county level random effects described by the spline function resulted from the estimated coefficients $\hat{\boldsymbol{\xi}}$. Clearly, the states of North Dakota, Iowa, Maryland, Virginia and southern region of Arizona have significantly fewer doctors than expected from their fixed effects. On the other hand, Nevada, the upper peninsula of Michigan, and Maine have areas with more doctors. Figure 2.3b shows the average

state-time random effect $\bar{\boldsymbol{s}}_{1i\cdot}$. We can see that the states with highly positive effects are Nevada, Idaho, South Dakota, and Maine. The states with highly negative effects are New Jersey, Louisiana, Minnesota, Montana, and Washington. Figure 2.3c shows the overall location effect for each county based on the sum of the two random effects ($\boldsymbol{S_1} + \boldsymbol{S_2}$). Nevada, Idaho, South Dakota, and Nebraska all show positive location effects, which may initially seem counter-intuitive given the small numbers shown in Figures 2.1a and 2.1b. However, these are less likely to have none or fewer primary care physicians relative to their county characteristics (population and other covariates). On the other hand, North Dakota, much of the mid-Atlantic coast, and the lower Mississippi River region have fewer physicians than expected. Finally, we also show the expected county level counts per 3500 residents (Figure 2.3d) which are calculated from the fixed effects and random effects specified in the models in equations (2.4) and (2.5). Comparing these with the observed counts in Figure 2.1a provides a population-standardized visual for the goodness of fit of the proposed model. We further evaluate the adequacy of our model performance using the posterior predictive fits, which are presented in Appendix A.2, indicating that the proposed model provides a good fit to the HPSA data.

2.4 Simulation study

We carried out simulation studies to examine the performance of our proposed model and its estimation method. For simplicity, we consider a regional subset of the country consisting of a rectangular collection of states covering portions of the Great Plains, Midwest, and mid-Atlantic regions. Specifically, we consider North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, Missouri, Wisconsin, Illinois, Michigan, Indiana, Ohio, Kentucky, New York, Pennsylvania, West Virginia, Maryland, and Virginia. We include 10 covariates as predictors: population size, median age, Asian(%), native American(%), urban influence, labor (%), per capita income, below

poverty line (%), births per 1000, deaths per 1000, as well as indicators for the year. The true data generating model is the negative binomial hurdle model introduced previously. Values for the coefficients were chosen to mimic the case study and can be found in Table 2.3. The true Ψ matrix was taken to have equicorrelation structure with correlation coefficient of 0.9 and a variance of 0.03. The knots for the spline basis functions were selected as before.

We generated data from the proposed hierarchical mixed effect hurdle model and fitted the data with the following four models:

1. **NB-SP-ST**: Negative binomial hurdle model with spline and state-time interaction, which is the full model described in Section 2.2.1.

2. **NB-ST**: Negative Binomial hurdle model with only state-time interaction, which is the full model without the county effect $\boldsymbol{S_2}$. That is the following model:

$$logit(p_{ijk}) = X_{ijk}^{'}\beta + a \cdot s_{1ik} \tag{2.11}$$

   and

$$log(\mu_{ijk}) = X_{ijk}^{'}\gamma + s_{1ik}. \tag{2.12}$$

3. **POI-SP-ST**: Poisson hurdle model with spline and state-time interaction, which is the same model as NB-SP-ST but with Poisson distribution assumption for counts. The model structure is the same as NB-SP-ST, however the dispersion parameter ϕ is set as 0.

4. **POI-ST**: Poisson hurdle model with only state-time interaction, which is the same model as NB-ST but with the dispersion parameter ϕ fixed as 0.

We generated 100 simulated data sets according to NB-SP-ST along with $\boldsymbol{S_1}$ and the coefficients $\boldsymbol{\xi}$ given as in the case study in Section 2.3.2. We consider the same estimation approach as detailed in Section 2.2.2. Note that when fitting NB-SP-ST

(or NB-SP), the first stage involves fitting POI-SP-ST (or POI-SP). To simplify, we consider only one run of the second stage for NB dispersion parameter ϕ. Let the estimated $\hat{\phi} = \phi_{(2)}$ be the maximizer of the log-likelihood function of equation (2.10) after the first fit to the negative binomial distribution. At each stage, we run 3 parallel chains for 500 iteration (discarding the first 200 iterations for burn-in), yielding 900 posterior samples for inference.

Results under the four models are presented in Table 2.3. We report the average estimate, bias, and mean squared error (MSE) for each parameter across the 100 replicated data sets. To summarize the accuracy within the model components, we consider a loss function for the binary and count models defined by the sum of squared errors for the fixed effect coefficients in the binary and count model specified in Section 2.2.1.

From the Table 2.3, it is evident that our model NB-SP-ST performed the best among the competitors, since the loss is minimized in our proposed model. Although NB-SP-ST and POI-SP-ST have the same model for the binary component, the estimates from the Negative Binomial model are better than those in the Poisson model, due to the shared random effects between the count and binary models. The loss of the count model under POI-SP-ST is 34% larger than that of under NB-SP-ST, and missing the spatial component increases the loss from NB-SP-ST to NB-ST by 68%. The true value of ϕ is covered by the 95% Wald-type credible interval in 98 out of the 100 simulated data sets, indicating appropriate coverage. In the NB-ST model, the coverage of the true ϕ is 0% since the absence of the spatial components in this model requires an inflated ϕ to account for this unexplained variation.

We additionally carried out simulations when the data is generated under the NB-ST and POI-SP-ST models, and the results are reported in Tables A1.1 and A1.2 in the Appendix A.3. Based on the simulation results, we conclude that the NB-SP-ST displays the best performance, even if the true data generating model is simpler.

In fact, NB-SP-ST has smaller loss than the true model in many cases.

2.5 Conclusion and discussion

In this work we have proposed a novel framework for modeling count data while accounting for excessive zeros and both temporal and spatial correlations. We accomplish this through a hurdle model that uses logistic regression for the zero component and a truncated negative binomial for the overdispersed positive counts. The complex correlation structure is accommodated by pairing a spline model to describe local/county-level deviations with a state-specific random effect that is correlated across the multiple years.

As mentioned, one of the key challenges in this work is computational. The use of `Stan` provides an accessible and automated software to sample from our model. However, computing is still somewhat slow. As mentioned previously, it is particularly problematic when the overdispersion ϕ is sampled. To avoid this issue, we develop an empirical Bayes approach that estimates the overdispersion $\hat{\phi}$ by maximizing the marginal posterior $\pi(\phi|\mathbf{Y})$. To represent the precision of this estimate, we propose a Wald-style credible interval (see Appendix A.1).

In the application of the HPSA data, we model the number of primary care physicians per county against a variety of representative covariates. As expected, population is the most important driver of medical coverage, but a variety of other demographic (racial and gender make-up), economic (income and poverty rate), and medical factors (e.g., presence of medical school and birth and death rates) also play important roles. Additionally, the estimated random effects and spline function can be used to assess local and state-level differences from those expected by the model. Even though both represent largely rural areas, South Dakota and Nebraska are shown to have more medical coverage than would be expected from their covariate characteristics, compared to Mississippi and Louisiana which have fewer primary care

physicians than expected. It would then be of interest to determine what factors (beyond the predictors of our model) or public policies may explain this difference and explore how these could be used to increase coverage in these under served areas.

2.6 Tables and Figures

Table 2.1: The variables which may impact the outcome variable (i.e. the number of primary care physicians)

Covariate name	Definition
Population size	Total population size for all ages in a county (in natural logarithmic scale)
Median age	Median age of the population of the county
African-American(%)	Percentage of the county population that is African-American
Asian(%)	Percentage of the county population that is Asian
Native American(%)	Percentage of the county population that is Native American
Hispanic(%)	Percentage of the county population that is Hispanic
Years of education	Median years of education of that county
Female(%)	Percentage of the county population that is female
18+ years education(%)	Percentage of the county population with 18+ years of education
Urban	Scale measuring how urban a county is
Labor(%)	Percentage of the county population participating in labor force
Unemployment rate(%)	Percentage of that county population that is unemployed
Per capita income	Per capita income of that county
Below poverty line(%)	Percentage of the population living below the poverty line
No health insurance(%)	Percentage of county population with no health insurance
65+ older(%)	Percentage of county population that is 65 or older
Medical school	Indicator of at least one hospital(s) with a medical school affiliation in the county
Annual doctor's wage	Mean annual doctor's wage in the state
State income tax rate	State income tax rate of the state
Births per 1000	Number of births per 1000 population in the county
Deaths per 1000	Number of deaths per 1000 population in the county
Gambling	Number of gambling facilities in the county
Boat marinas	Number of boat marinas in the county

Table 2.2: Fixed and random effect parameter estimates and their 95 % credible intervals for binary and count models

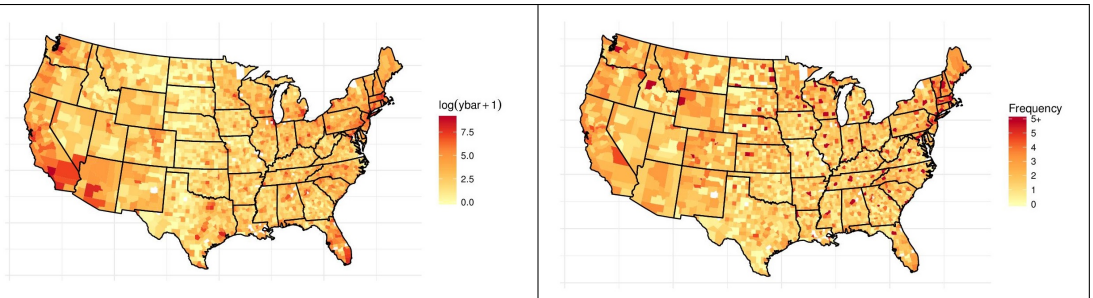
	Covariates	Binary Model			Count Model		
		Est (β)	95 % CI		Est (γ)	95 % CI	
Fixed effects	Intercept	5.684	5.278	6.166	2.614	2.552	2.679
	Population size(log)	**3.913**	**3.665**	**4.162**	**1.687**	**1.672**	**1.701**
	Median Age	0.090	-0.076	0.262	**-0.030**	**-0.047**	**-0.015**
	African American(%)	**0.147**	**0.040**	**0.262**	**0.053**	**0.040**	**0.067**
	Asian(%)	**-0.597**	**-0.847**	**-0.310**	**-0.039**	**-0.047**	**-0.030**
	Native American(%)	-0.037	-0.115	0.039	-0.008	-0.020	0.004
	Hispanic(%)	**0.127**	**0.022**	**0.228**	**0.044**	**0.028**	**0.060**
	Years of education	**-0.169**	**-0.307**	**-0.032**	0.000	-0.010	0.010
	Female(%)	**0.107**	**0.045**	**0.169**	**0.025**	**0.015**	**0.036**
	18+ years education(%)	**0.777**	**0.606**	**0.953**	**0.284**	**0.272**	**0.297**
	Urban	**0.318**	**0.214**	**0.420**	**0.097**	**0.086**	**0.110**
	Labor(%)	0.089	-0.006	0.186	**0.076**	**0.063**	**0.089**
	Unemployment rate	-0.124	-0.249	0.003	-0.014	-0.028	0.000
	Per capita income	**0.164**	**0.042**	**0.288**	**0.048**	**0.035**	**0.061**
	Below poverty line(%)	**-0.183**	**-0.313**	**-0.045**	**0.028**	**0.013**	**0.043**
	No health insurance(%)	**0.124**	**0.016**	**0.236**	**-0.079**	**-0.096**	**-0.062**
	65+ older(%)	**0.236**	**0.058**	**0.419**	**0.052**	**0.032**	**0.073**
	Medical school	**1.013**	**0.418**	**1.600**	**0.242**	**0.225**	**0.260**
	Annual doctor's wage	0.092	-0.029	0.223	-0.001	-0.022	0.021
	State income tax	**0.253**	**0.089**	**0.419**	0.014	-0.028	0.056
	Births per 1000	**0.357**	**0.244**	**0.465**	**0.098**	**0.085**	**0.111**
	Deaths per 1000	**0.207**	**0.120**	**0.298**	**0.169**	**0.152**	**0.186**
	Gambling	0.139	-0.137	0.576	**-0.010**	**-0.016**	**-0.005**
	Boat marinas	0.635	-0.582	3.031	**-0.021**	**-0.027**	**-0.015**
	Year 2007	0.251	-0.157	0.675	0.046	-0.006	0.099
	Year 2008	0.208	-0.135	0.571	0.012	-0.034	0.058
	Year 2009	0.048	-0.255	0.368	0.018	-0.023	0.058
	Year 2010	0.210	-0.114	0.526	0.043	0.004	0.081
	Year 2011	0.045	-0.232	0.328	0.016	-0.022	0.056
	Year 2012 (ref)	0	-	-	0	-	-
Random*	a	3.129	2.519	3.804	-	-	-
	σ_ξ	-	-	-	0.449	0.390	0.512
	ϕ	-	-	-	0.0870	0.0841	0.0900

* indicates random components

Table 2.3: Simulation results when the underlying model was NB-SP-ST, but the data was fitted with different models

		True	**NB-SP-ST**			**NB-ST**			**POI-SP-ST**			**POI-ST**		
		Value	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$
Binary Model	β_0	2.2	2.185	-0.015	3.084	2.109	-0.091	5.698	2.177	-0.023	3.475	2.085	-0.115	6.177
	β_1	1.3	1.308	0.008	0.531	1.234	-0.066	1.284	1.302	0.002	0.585	1.225	-0.075	1.475
	β_2	0	0.011	0.011	0.345	0.010	0.010	0.619	0.011	0.011	0.328	0.008	0.008	0.632
	β_3	-0.1	-0.079	0.021	0.984	-0.051	0.049	1.433	-0.078	0.022	1.003	-0.050	0.050	1.505
	β_4	0	-0.001	-0.001	0.114	-0.006	-0.006	0.322	-0.002	-0.002	0.118	-0.008	-0.008	0.365
	β_5	0.1	0.099	-0.001	0.293	0.086	-0.014	0.506	0.098	-0.002	0.308	0.085	-0.015	0.557
	β_6	0	-0.005	-0.005	0.283	-0.009	-0.009	0.521	-0.005	-0.005	0.298	-0.007	-0.007	0.555
	β_7	0	-0.001	-0.001	0.281	-0.007	-0.007	0.491	-0.002	-0.002	0.299	-0.011	-0.011	0.515
	β_8	0	-0.001	-0.001	0.374	-0.003	-0.003	0.842	0.001	0.001	0.383	-0.002	-0.002	0.898
	β_9	0.1	0.106	0.006	0.221	0.105	0.005	0.336	0.106	0.006	0.220	0.107	0.007	0.367
	β_{10}	0.1	0.100	0.000	0.192	0.089	-0.011	0.350	0.099	-0.001	0.190	0.090	-0.010	0.373
	$t1_1$	0.05	0.048	-0.002	1.514	0.040	-0.010	1.811	0.047	-0.003	1.414	0.035	-0.015	1.785
	$t1_2$	0.1	0.101	0.001	1.781	0.093	-0.007	1.670	0.104	0.004	1.855	0.092	-0.008	1.764
	$t1_3$	0.15	0.169	0.019	1.803	0.162	0.012	1.767	0.173	0.023	1.865	0.159	0.009	1.818
	$t1_4$	0.2	0.186	-0.014	1.598	0.176	-0.024	1.608	0.182	-0.018	1.597	0.173	-0.027	1.603
	$t1_5$	0.25	0.253	0.003	1.482	0.244	-0.006	1.562	0.249	-0.001	1.538	0.237	-0.013	1.552
	Loss*				**14.879**			20.819			15.478			21.940
Count Model	γ_0	2.2	2.190	-0.010	0.234	2.215	0.015	0.437	2.193	-0.007	0.290	2.215	0.015	0.433
	γ_1	1.5	1.502	0.002	0.012	1.500	0.000	0.023	1.500	0.000	0.022	1.496	-0.004	0.055
	γ_2	0	0.000	0.000	0.013	0.002	0.002	0.032	0.000	0.000	0.021	0.000	0.000	0.055
	γ_3	-0.9	-0.901	-0.001	0.021	-0.901	-0.001	0.050	-0.900	0.000	0.034	-0.899	0.001	0.100
	γ_4	0	0.000	0.000	0.010	0.000	0.000	0.025	0.000	0.000	0.013	-0.002	-0.002	0.033
	γ_5	0.2	0.202	0.002	0.007	0.201	0.001	0.026	0.201	0.001	0.011	0.199	-0.001	0.056
	γ_6	0	0.001	0.001	0.010	0.002	0.002	0.035	0.002	0.002	0.017	0.004	0.004	0.061
	γ_7	0.1	0.100	0.000	0.012	0.101	0.001	0.032	0.100	0.000	0.024	0.101	0.001	0.056
	γ_8	-0.3	-0.300	0.000	0.014	-0.302	-0.002	0.045	-0.299	0.001	0.023	-0.299	0.001	0.079
	γ_9	0.2	0.200	0.000	0.008	0.199	-0.001	0.016	0.199	-0.001	0.015	0.201	0.001	0.047
	γ_{10}	0.2	0.200	0.000	0.010	0.201	0.001	0.027	0.200	0.000	0.014	0.200	0.000	0.038
	$t2_1$	0.025	0.023	-0.002	0.067	0.023	-0.002	0.100	0.024	-0.001	0.096	0.023	-0.002	0.126
	$t2_2$	0.05	0.049	-0.001	0.083	0.048	-0.002	0.113	0.050	0.000	0.109	0.048	-0.002	0.157
	$t2_3$	0.075	0.071	-0.004	0.088	0.070	-0.005	0.105	0.073	-0.002	0.102	0.072	-0.003	0.128
	$t2_4$	0.1	0.099	-0.001	0.070	0.098	-0.002	0.082	0.100	0.000	0.098	0.100	0.000	0.113
	$t2_5$	0.125	0.123	-0.002	0.067	0.121	-0.004	0.071	0.121	-0.004	0.087	0.120	-0.005	0.097
	Loss*				**0.727**			1.219			0.978			1.634
	ϕ	0.096	0.097	0.001	**0.001**	0.121	0.025	0.066	-	-	-	-	-	-
	a	3.5	3.457	-0.043	**3.412**	3.263	-0.237	11.976	3.255	-0.245	9.741	2.925	-0.575	41.804

Loss*: The sum of squared errors for the fixed effect coefficients



(a) County level physicians on log-scale

(b) County level physicians per 3500 persons

Figure 2.1: Observed number of primary care physicians at county level (on the log-scale) is shown in panel (a), and the observed number of primary care physicians per 3500 persons is shown in panel (b). If the observed number of primary care physicians per 3500 persons is 6 or more, we label it as "5+".

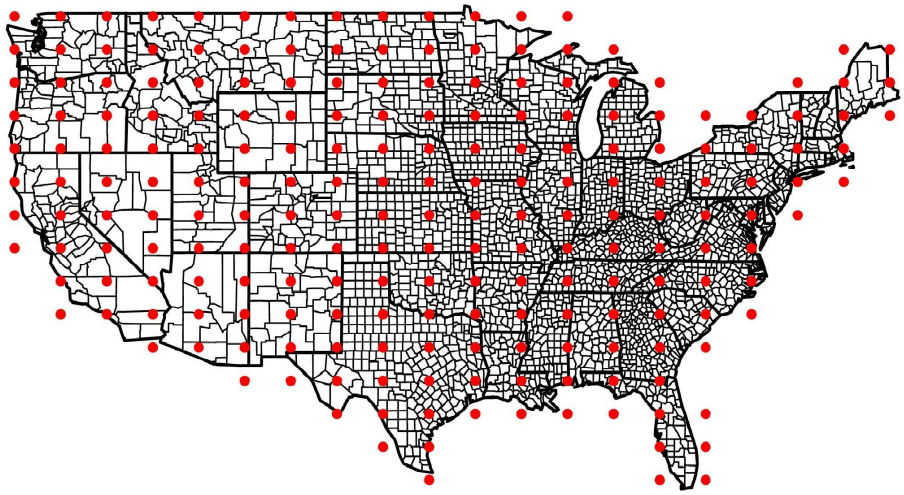
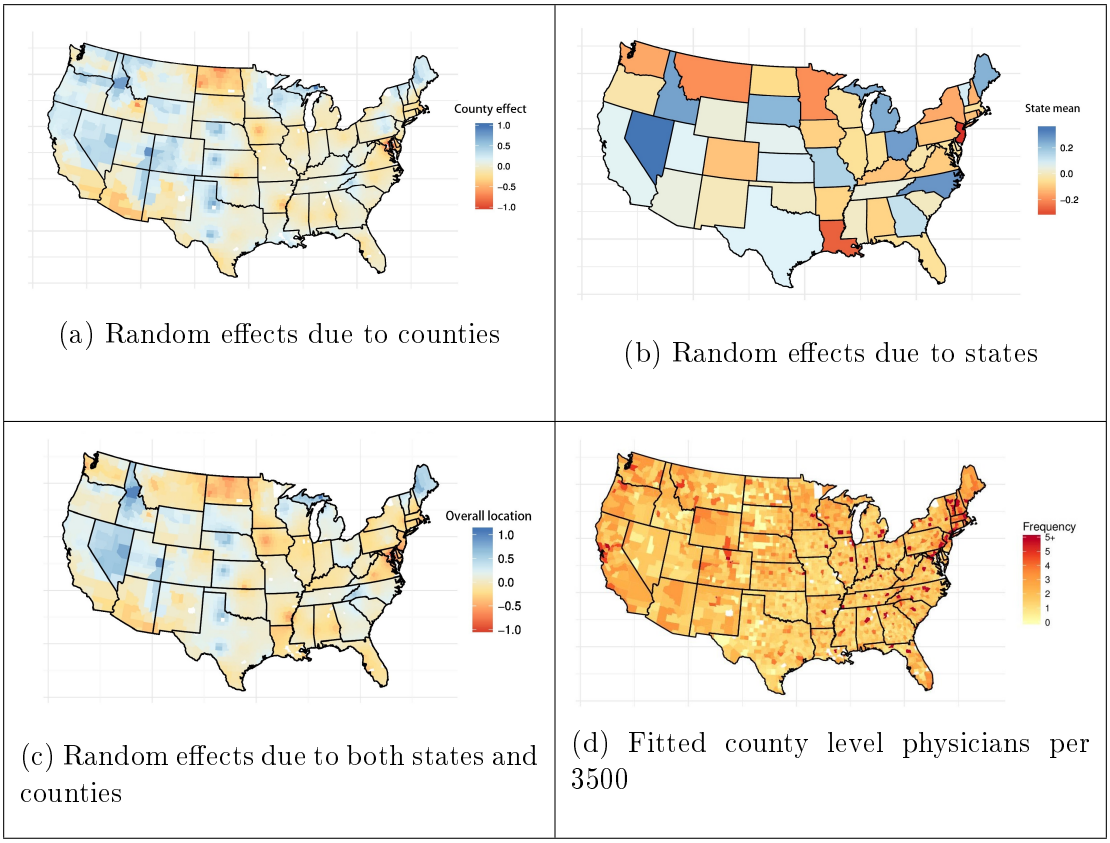


Figure 2.2: The illustration of the selected knots



(a) Random effects due to counties

(b) Random effects due to states

(c) Random effects due to both states and counties

(d) Fitted county level physicians per 3500

Figure 2.3: Illustration of different location effects

CHAPTER 3

COMPARISONS OF AVERAGE TREATMENT EFFECTS FOR MULTIPLE GROUPS WHEN OUTCOME IS ORDINAL AND CONFOUNDING EXISTS

3.1 Introduction

Ordinal outcome variables are very common in clinical studies. Accurately assessing the treatment effects (ATE) from two or more treatment groups is important for treatment selection and patients care. Randomized controlled trials (RCT) are considered as gold standards to estimate ATE since RCT often implies the unconfoundedness (i.e., both observed and unobserved confounding are independent of the treatment assignment) (Austin, 2011). The unconfoundedness also implies that all the covariates (observed or unobserved) are balanced among different groups. The popular parametric methods for ordinal outcome variables are ordinal cumulative link models (Agresti, 2007, 2010; Ryu and Agresti, 2008), where the probability of the outcome variable being less than a certain level is linked with the linear combination of covariates via a cumulative distribution function (Agresti and Kateri, 2017). Two commonly used ordinal cumulative link models are (1) the ordinal logistic regression model where the link function is the cumulative logistic function and (2) the cumulative probit model, where the link function is the cumulative normal distribution. The most popular non-parametric method to compare two treatment groups for the ordinal outcome is the Mann-Whitney U-statistic (Mann and Whitney, 1947), which

assumes no confounding covariates. However RCT may not be always feasible due to ethics, cost and patient preferences (Kang and Schafer, 2007; Robins et al., 1995). On the other hand, with the availability of the observed data in natural health care setting, estimating the treatment effect based on observational studies becomes more practical. In the observational studies, the confounding covariates often exist, and the statistical methods developed for RCT may not be suitable any more.

Rosenbaum and Rubin (1983) introduced propensity score based methods to estimate treatment effects between two groups based on observational data. The propensity score technique mimics randomization procedure to balance covariates when randomization is not made. Propensity score is defined as the conditional probability of receiving a treatment given all the covariates. The fundamental idea for the propensity score based approach is that the response is independent of the treatment assignment given the propensity score. Rosenbaum and Rubin (1983, 1985) showed that propensity score based approach could remove all bias of the ATE estimate. In randomized experiments, the true propensity scores are known and hence the propensity score based methods are very efficient (Austin, 2011; Hirano et al., 2003). In non-randomized experiments, propensity scores can be estimated using different techniques: parametric model such as a logistic regression model or nonparametric models such as random forests, bagging (Lee et al., 2010; McCaffrey et al., 2004), and generalized boosting methods (Abdia et al., 2017). Once propensity score is estimated, various propensity score based methods such as matching (Heckman et al., 1998; Rosenbaum and Rubin, 1983), stratification (Rosenbaum, 1987; Rosenbaum and Rubin, 1983) and inverse-probability of treatment weighting (IPW) (Austin and Stuart, 2015; Hirano et al., 2003; Horvitz and Thompson, 1952; Robins et al., 1995) have been used to estimate the average treatment effects.

When there are multiple treatment groups, Imbens (2000) proposed using generalized propensity score (GPS) to balance covariates and control confounding vari-

ables. GPS can be estimated using parametric methods, such as multinomial logistic regression model (Satten et al., 2018) or ordinal logistic regression model (Robins, 2000). GPS has also been estimated nonparametrically using machine learning methods such as bagging and generalized boosting methods (Abdia et al., 2017).

We investigate the GPS based ordinal cumulative link models and GPS adjusted U-statistics to compare treatment effects among multiple groups. We focus on parametric approach to estimate GPS so that the asymptotic variance for the adjusted U-statistics can be constructed. GPS-based regression and stratification have been investigated in estimating ATE when outcome is continuous variable. The investigation for GPS-based methods for estimating ATE is limited when the outcome is ordinal variable. Some methods suitable for continuous outcome may not be suitable any more for ordinal outcome. In this project, we use superiority score (Agresti and Kateri, 2017) as a measure of treatment effect between two groups. We compare the performance of parametric approach and GPS-based approaches by using extensive simulations. Pros and cons are provided for each method. A case study is provided to study the effect of cadmium and arsenic on chronic kidney disease based on the National Health and Nutrition Examination Survey (NHANES) 2011-2014 data set.

3.2 Data structure and cumulative link models

Let $\mathbf{Y}$ denote the ordinal outcome variable with $\mathcal{C}$ categories (*i.e.*, $\mathbf{Y} \in \{1, \ldots, \mathcal{C}\}$). Let assume that there are N subjects in the sample. $\mathbf{X}_i, Y_i$, and T_i $(i = 1, \ldots, N)$ denote respectively a vector of p confounding variables, the ordinal outcome variable, and the treatment status for the i^{th} subject in the study sample. Let us assume that there are $\mathcal{M}$ treatment groups, and $T_i = t$ if i^{th} subject gets the t^{th} treatment, where $t \in \{0, 1, \ldots, \mathcal{M} - 1\}$. Using the potential outcome notations (Rubin, 1974)

for $\mathcal{M} = 2$, we can write the observed outcome Y_i as

$$Y_i = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)} = \begin{cases} Y_i^{(0)}, & \text{if } T_i = 0, \\ Y_i^{(1)}, & \text{if } T_i = 1. \end{cases} \tag{3.1}$$

Here $Y_i^{(0)}$ is the potential outcome when the i^{th} subject is in control group, and $Y_i^{(1)}$ is the potential outcome when the i^{th} subject is in the treatment group. Both $Y_i^{(0)}$ and $Y_i^{(1)}$ are $\mathcal{C}$-category ordinal variable. However, only one of the potential outcomes is observed depending on the treatment received. For $\mathcal{M} > 2$, using the potential outcome notation the observed Y_i can be written as

$$Y_i = \sum_{t=0}^{\mathcal{M}-1} Y_i^{(t)} I\{T_i = t\} = \begin{cases} Y_i^{(0)}, & \text{if } T_i = 0, \\ Y_i^{(1)}, & \text{if } T_i = 1, \\ \dots & \\ Y_i^{(\mathcal{M}-1)}, & \text{if } T_i = \mathcal{M} - 1. \end{cases} \tag{3.2}$$

Here $Y_i^{(t)}$ denotes the potential outcome if i^{th} subject were treated with t^{th} treatment $(t = 0, 1, \cdots, \mathcal{M} - 1)$. However, we assume that one subject only receives one treatment which corresponds the treatment assigned. Let n_t denotes the number of observations from group t, Y_i is the observed outcome for the i^{th} subject, we have $N(= n_0 + n_1 + \ldots + n_{\mathcal{M}-1})$ observations in total. In general, to test the treatment effects for $\mathcal{M}$ groups we can construct $\mathcal{M} - 1$ contrasts and test the $\mathcal{M} - 1$ contrasts simultaneously.

3.2.1 Estimand for ATE for ordinal outcome

When outcome variable is ordinal, superiority score has been proposed as a measure of treatment effect for two groups (Agresti and Kateri, 2017). Let assume $Y^{(0)}$ and $Y^{(1)}$ as two potential outcomes under control and treatment group respectively. The

stochastic superiority score of $Y^{(1)}$ over $Y^{(0)}$ (Klotz, 1966; Kruskal, 1957; Vargha and Delaney, 1998) is defined as

$$\gamma = P(Y^{(0)} < Y^{(1)}) + \frac{1}{2}P(Y^{(0)} = Y^{(1)}), \tag{3.3}$$

where the term $P(Y^{(0)} = Y^{(1)})$ adjusts for ties. Note that the form of γ in (3.3) is closely related to the Wilcoxon kernel (Wilcoxon, 1945). If the outcome variable is continuous, then $P[Y^{(0)} = Y^{(1)}] = 0$, resulting in $\gamma = P(Y^{(0)} < Y^{(1)})$.

Under the null hypothesis that there is no treatment effect, we have $\gamma = \frac{1}{2}$. The hypothesis test on whether there is a treatment effect for two groups is equivalent to test $H_0 : \gamma = \frac{1}{2}$ versus $H_1 : \gamma \neq \frac{1}{2}$. $\gamma > \frac{1}{2}$ indicates that $Y^{(1)}$ is stochastically superior than $Y^{(0)}$. To estimate the superiority score and test for the hypotheses, both parametric and non-parametric methods are proposed.

3.2.2 Parametric approaches to estimate ATE

In this section, we introduce the parametric models for ordinal outcomes using the cumulative link models (Agresti and Kateri, 2017). The cumulative link models can be expressed via a continuous latent variable. Let Y denote an ordinal response variable with $\mathcal{C}$ possible outcomes, and Y^* a continuous latent variable associated with Y. Y could be defined via the latent variable Y^* and the $\mathcal{C}-1$ cut points $(\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{C}-1})$ with $\alpha_1 < \alpha_2 < \ldots < \alpha_{\mathcal{C}-1}$. That is,

$$Y = \begin{cases} 1 & \text{if } Y^* \leq \alpha_1; \\ 2 & \text{if } \alpha_1 < Y^* \leq \alpha_2; \\ \ldots & \\ \mathcal{C}-1 & \text{if } \alpha_{\mathcal{C}-2} < Y^* \leq \alpha_{\mathcal{C}-1}; \\ \mathcal{C} & \text{if } Y^* > \alpha_{\mathcal{C}-1}. \end{cases} \tag{3.4}$$

Agresti and Kateri (2017) assumed that the latent variable Y^* follows a certain continuous distribution with mean $\tau T + X^{'}\boldsymbol{\delta}_x$ and variance 1, where $X = (X_1, \ldots, X_p)^{'}$ are the explanatory variables with regression coefficients $\boldsymbol{\delta}_x = (\delta_1, \ldots, \delta_p)^{'}$, T is a binary variable indicating whether the subject is in treatment or control group. $(Y^* - \tau T - X^{'}\boldsymbol{\delta}_x)$ has a cumulative distribution function (CDF) $\mathcal{F}(\cdot)$ with mean 0 and variance 1. From equation (3.4) we have,

$$\begin{aligned} P[Y \leq j|X,T] &= P[Y^* \leq \alpha_j|X,T] \\ &= P[Y^* - \tau T - X^{'}\boldsymbol{\delta}_x \leq \alpha_j - \tau T - X^{'}\boldsymbol{\delta}_x] \\ &= \mathcal{F}(\alpha_j - \tau T - X^{'}\boldsymbol{\delta}_x). \end{aligned}$$

Hence, the cumulative link models can be written as

$$\mathcal{F}^{-1}(P[Y \leq j|X,T]) = \alpha_j - \tau T - X^{'}\boldsymbol{\delta}_x, \quad j = 1, \ldots, \mathcal{C} - 1. \tag{3.5}$$

Theoretically, the link function $\mathcal{F}^{-1}$ could be the inverse function of any CDF of a continuous variable. The commonly used link functions are probit and logit link functions. The probit link function $\mathcal{F}$ is taken as the CDF of a standard normal distribution (*i.e.*, $\mathcal{F} = \Phi$). With probit link function, equation (3.5) becomes

$$\Phi^{-1}(P[Y \leq j|X,T]) = \alpha_j - \tau T - X^{'}\boldsymbol{\delta}_x, \quad j = 1, \ldots, \mathcal{C} - 1.$$

Hence for probit model, the superiority score, say γ_{probit}, has the form

$$\begin{aligned} \gamma_{probit} = P(Y^{(0)*} < Y^{(1)*}|X,T) &= P\left[\frac{Y^{(1)*} - Y^{(0)*} - \tau}{\sqrt{2}} > \frac{-\tau}{\sqrt{2}}\right] \\ &= \Phi\left(\frac{\tau}{\sqrt{2}}\right). \end{aligned} \tag{3.6}$$

Here we assume that $Y^{(0)*} \sim N\left(X^{'}\boldsymbol{\delta}_x, 1\right)$ and $Y^{(1)*} \sim N\left(\tau + X^{'}\boldsymbol{\delta}_x, 1\right)$, which implies

that $Y^{(1)*} - Y^{(0)*} \sim N(\tau, 2)$. The parameters $(\alpha_1, \ldots, \alpha_{\mathcal{C}-1}, \boldsymbol{\delta}_x, \tau)$ can be estimated from the maximum likelihood estimations, and the superiority score is estimated as

$$\hat{\gamma}_{probit} = \Phi\left(\frac{\hat{\tau}}{\sqrt{2}}\right). \tag{3.7}$$

If we take the link function as logit link function, i.e., $\mathcal{F}(x) = \frac{e^x}{1+e^x}$, then equation (3.5) becomes the ordinal logistic regression model:

$$log\left(\frac{P[Y \leq j|X,T]}{1 - P[Y \leq j|X,T]}\right) = \alpha_j - \tau T - X^{'}\boldsymbol{\delta}_x, \quad j = 1, \ldots, \mathcal{C}-1.$$

Under logistic regression model, the superiority score doesn't have a closed from. Agresti and Kateri (2017) provided an approximate estimation for the superiority score:

$$\hat{\gamma}_{logit} \approx \frac{exp(\hat{\tau}/\sqrt{2})}{1 + exp(\hat{\tau}/\sqrt{2})}. \tag{3.8}$$

The cumulative link models could be extended to examine the treatment effect among multiple treatment groups by using the following form:

$$\mathcal{F}^{-1}(P[Y \leq j|X,T]) = \alpha_j - \tau_1 T^{(1)} - \ldots - \tau_{\mathcal{M}-1} T^{(\mathcal{M}-1)} - X^{'}\boldsymbol{\delta}_x, \quad j = 1, \ldots, \mathcal{C}-1. \tag{3.9}$$

Here $T^{(1)}, \ldots, T^{(\mathcal{M}-1)}$ are dummy variables to indicate the treatment assignment groups. $T^{(t)} = 1$ if the subject comes from treatment group t, and $T^{(t)} = 0$ otherwise $(t \in \{1, 2, \ldots, \mathcal{M}-1\})$. Here we set control group as reference group. The other options for specifying $\mathcal{M}-1$ variables to the $\mathcal{M}$ treatment groups are also feasible, which are not detailed here. Under the model (3.9), testing whether there is any treatment effect is equivalent as testing $H_0 : \tau_1 = \ldots = \tau_{\mathcal{M}-1} = 0$ vs H_a : at least one τ_t is not 0. The comparison between group t and $t^{'}$ can be made via comparing τ_t and $\tau_{t'}$. Here τ_0 is set as 0. Although the complementary log-log link function can also be applied to model the ordinal outcome, we focus only on cumulative logit model and probit

model for their popularity.

3.3 Proposed GPS-based approaches

3.3.1 Generalized propensity score (GPS) models

In multiple treatment setting, generalized propensity score (GPS) could be estimated using parameter methods and machine learning methods. Here we focus on parametric methods. One of the popular parametric methods is using multinomial logistic regression (Imbens, 2000; Satten et al., 2018):

$$\log\left\{\frac{P[T=t|X=x]}{P[T=0|X=x]}\right\} = X^{'}\beta^{(t)}, \text{ for } t=1,\ldots,\mathcal{M}-1. \tag{3.10}$$

Here the parameters involved are $\boldsymbol{\beta} = \left(\beta^{(1)},\ldots,\beta^{(\mathcal{M}-1)}\right)$, which includes $(\mathcal{M}-1)\times(p+1)$ parameters.

We also adopt the ordinal logistic regression (see Robins (2000)) to estimate GPS:

$$\log\left\{\frac{P[T<t|X=x]}{P[T\geq t|X=x]}\right\} = \beta_{0t} + X^{'}\beta, \text{ for } t=1,\ldots,\mathcal{M}-1. \tag{3.11}$$

Here the parameters are $\boldsymbol{\beta} = (\beta_{01},\ldots,\beta_{0\mathcal{M}-1},\beta)$, which includes $(\mathcal{M}-1+p)$ parameters. The GPS specified in equation (3.11) is more attainable when the number of treatment group is large. This specification in (3.11) also makes it feasible to stratify the subjects into different strata so that the superiority score between any two groups within the same strata can be estimated. $\boldsymbol{\beta}$ in model (3.10) and model (3.11) can be obtained by solving the estimating equations of the form

$$\sum_{i=1}^{n} \boldsymbol{S}_i(\boldsymbol{\beta}) = 0, \tag{3.12}$$

where $\boldsymbol{S}_i(\boldsymbol{\beta})$ is the score function of $\boldsymbol{\beta}$ from the i^{th} observation. Define $\hat{\boldsymbol{S}}_i = \boldsymbol{S}_i(\hat{\boldsymbol{\beta}})$

and $\hat{\boldsymbol{J}} = \sum_{i=1}^{N} \frac{\partial \boldsymbol{S}_i}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$. The forms of $\boldsymbol{S}_i(\boldsymbol{\beta})$, when treatment is generated from multinomial logistic regression model (3.10) or ordinal logistic regression model (3.11), are provided in Appendix A.5.

3.3.2 Adjusted U-Statistic for ordinal outcome

U-statistics are used quite often to compare the distributions of response among two or multiple treatment groups (Kruskal, 1957; Mann and Whitney, 1947). The most simplified U-statistic to compare distributions of two groups (e.g., control versus treatment group 1) is given by Mann and Whitney (1947):

$$U = \frac{1}{n_0}\frac{1}{n_1} \sum_{i \in \{i:T_i=0\}} \sum_{j \in \{j:T_j=1\}} K(Y_i, Y_j). \tag{3.13}$$

where $K(Y_i, Y_j) = I(Y_i < Y_j) + \frac{1}{2} I(Y_i = Y_j)$, for estimating the superiority score. U-statistics are very general and can be used for ordinal categorical non-numeric response. However the classic U-statistics can't account for confounding covariates, which may lead to large biased estimate for treatment effect. Satten et al. (2018) proposed adjusted U-statistics to estimate association or treatment effect with GPS estimated by multinomial logistic regression model. However, the performance of U-statistic in estimating superiority score for ordinal outcome has not been evaluated in particular, when the GPS is estimated by ordinal logistic regression model. In the following, we construct the inverse probability weighted (IPW) U-statistics and derive their asymptotic variances, where GPS can be estimated using either multinomial regression model or ordinal logit model.

Once GPS is estimated, IPW technique is applied to weight each observation into the equivalent size in the study sample. That is, the weight of the i^{th} subject

$w(x_i, t_i; \boldsymbol{\beta})$ is defined as

$$w(x_i, t_i; \boldsymbol{\beta}) = \frac{1}{P[T = t_i | X = x_i; \boldsymbol{\beta}]}, \ \ (i = 1, \cdots, N). \tag{3.14}$$

Thus, the two-sample (e.g., group t versus group $t^{'}$) adjusted U-statistics can be written as

$$U_a^{(tt^{'})} = \frac{1}{\sum_{i:T_i=t} w(x_i, t_i; \hat{\boldsymbol{\beta}})} \frac{1}{\sum_{j:T_j=t^{'}} w(x_j, t_j; \hat{\boldsymbol{\beta}})} \sum_{i:T_i=t} \sum_{j:T_j=t^{'}} K(y_i, y_j) w(x_i, t_i; \hat{\boldsymbol{\beta}}) w(x_j, t_j; \hat{\boldsymbol{\beta}}), \tag{3.15}$$

where $\sum_{i:T_i=t} w(x_i, t_i; \hat{\boldsymbol{\beta}})$ and $\sum_{j:T_j=t^{'}} w(x_j, t_j; \hat{\boldsymbol{\beta}})$, respectively, could be considered as the pseudo sample size (Robins, 2000) of group t and $t^{'}$ in the entire sample. Let denote,

$$\tilde{w}_t(x_i, t_i; \hat{\boldsymbol{\beta}}) = \frac{w(x_i, t_i; \hat{\boldsymbol{\beta}})}{W_t(\hat{\boldsymbol{\beta}})} \tag{3.16}$$

$$\text{and } \tilde{w}_{t^{'}}(x_j, t_j; \hat{\boldsymbol{\beta}}) = \frac{w(x_j, t_j; \hat{\boldsymbol{\beta}})}{W_{t^{'}}(\hat{\boldsymbol{\beta}})}. \tag{3.17}$$

Here $W_t(\hat{\beta}) = \frac{1}{n_t} \sum_{i:T_i=t} w(x_i, t_i; \hat{\boldsymbol{\beta}})$ and $W_{t^{'}}(\hat{\beta}) = \frac{1}{n_{t^{'}}} \sum_{j:T_j=t^{'}} w(x_j, t_j; \hat{\boldsymbol{\beta}})$. Then the adjusted U-statistics in equation (3.18) is equivalent to:

$$U_a^{(tt^{'})} = \frac{1}{n_t} \frac{1}{n_{t^{'}}} \sum_{\{i:T_i=t\}} \sum_{\{j:T_j=t^{'}\}} \big\{\widetilde{w}_t(x_i, t_i; \hat{\boldsymbol{\beta}})\big\} K(y_i, y_j) \big\{\widetilde{w}_{t^{'}}(x_j, t_j; \hat{\boldsymbol{\beta}})\big\}. \tag{3.18}$$

Equation (3.18) can be considered as the weighted kernel (Satten et al., 2018). Although Satten et al. (2018) laid the general framework for weighted U-statistics, we provide the novel application of this general work to estimate treatment effects of multiple groups for ordinal outcome.

Let assume $\mu^{(tt^{'})} = E\left(U_a^{(tt^{'})}\right)$. By following the derivation presented in Satten

et al. (2018) and using the Hoeffding decomposition (Hoeffding, 1948)

$$U_a^{(tt')} - \mu^{(tt')} \approx \sum_{i:T_i=t} \xi_t(i) + \sum_{j:T_j=t'} \xi_{t'}(j). \tag{3.19}$$

Here,

$$\begin{aligned}
\xi_t(i) &= \frac{1}{n_t}\left[-\mu^{(tt')}\left(\tilde{w}_t(X_i,t_i;\boldsymbol{\beta})-1\right)+\left(\tilde{h}_t(i)-\mu^{(tt')}\right)\right]+\frac{1}{n}C_nJ^{-1}S_i(X_i,t_i;\boldsymbol{\beta}),\\
\xi_{t'}(j) &= \frac{1}{n_{t'}}\left[-\mu^{(tt')}\left(\tilde{w}_{t'}(X_j,t_j;\boldsymbol{\beta})-1\right)+\left(\tilde{h}_{t'}(j)-\mu^{(tt')}\right)\right]+\frac{1}{n}C_nJ^{-1}S_j(X_j,t_j;\boldsymbol{\beta}),\\
\tilde{h}_t(i) &= \frac{1}{n_{t'}}\sum_{\{j:T_j=t'\}}\left\{\widetilde{w}_t(x_i,t_i;\boldsymbol{\beta})\right\}K(y_i,y_j)\left\{\widetilde{w}_{t'}(x_j,t_j;\boldsymbol{\beta})\right\},\\
\tilde{h}_{t'}(j) &= \frac{1}{n_t}\sum_{\{i:T_i=t\}}\left\{\widetilde{w}_t(x_i,t_i;\boldsymbol{\beta})\right\}K(y_i,y_j)\left\{\widetilde{w}_{t'}(x_j,t_j;\boldsymbol{\beta})\right\},\\
J &= plim\ n^{-1}\sum_{i=1}^{n}\frac{\partial S_i(x_i,t_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \approx -\left[\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\\
C_n &= \frac{\mu^{(tt')}}{n_t}\sum_{i:T_i=t}\widetilde{w}_t(x_i,t_i;\boldsymbol{\beta})\frac{\partial\log w(x_i,t_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}+\frac{\mu^{(tt')}}{n_{t'}}\sum_{j:T_j=t'}\widetilde{w}_{t'}(x_j,t_j;\boldsymbol{\beta})\frac{\partial\log w(x_j,t_j;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\\
&\quad+\frac{1}{n_tn_{t'}}\sum_{i:T_i=t}\sum_{j:T_j=t'}\left\{\widetilde{w}_t(x_i,t_i;\boldsymbol{\beta})\right\}K(y_i,y_j)\left\{\widetilde{w}_{t'}(x_j,t_j;\boldsymbol{\beta})\right\}\omega_{\boldsymbol{\beta}}(x_i,x_j,t_i,t_j;\boldsymbol{\beta})
\end{aligned}$$

where $\omega_{\boldsymbol{\beta}}(x_i,x_j,t_i,t_j;\boldsymbol{\beta}) = \left(\frac{\partial\log w(x_i,t_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}+\frac{\partial\log w(x_j,t_j;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)$. In particular, when GPS is estimated by multinomial regression in (3.10),

$$\frac{\partial\log w(x_i,t_i;\boldsymbol{\beta})}{\partial\beta_t} = \left(P\left[T=t|X=X_i,\boldsymbol{\beta}\right]-I_{\{t_i=t\}}\right)X_i,\ (t=1,\ldots,\mathcal{M}-1)$$

. When GPS is estimated by ordinal logistic regression in (3.11), the form of $\frac{\partial\log w(x_i,t_i;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$ is given in Appendix A.5. Satten et al. (2018) showed that $U_a^{(tt')}$ has an asymptotically normal distribution with mean $w=\frac{1}{2}$ and asymptotically variance given as,

$$\hat{v} = n_t\hat{\sigma}_t^2 + n_{t'}\hat{\sigma}_{t'}^2. \tag{3.20}$$

Here, ${\hat{\sigma_t}}^2 = \frac{1}{n_t - 1} \sum_{i:T_i=t} \left(\xi_t(i) - \bar{\xi_t}\right)^2$ and $\bar{\xi}_t = \dfrac{1}{n_t} \sum_{i:T_i=t} \xi_t(i)$.

To examine the treatment effect between group t and group $t^{'}$, we can estimate the superiority score and construct its 95% confidence interval. To test the null hypothesis that there is no difference between group t and group $t^{'}$, we can construct the test statistic by using the following asymptotic property under the null hypothesis:

$$Z_a = \frac{U_a - \frac{1}{2}}{\sqrt{\hat{v}}} \sim N(0, 1).$$

To test the null hypothesis that the distribution of Y from different groups are the same, we may construct $\mathcal{M} - 1$ linearly independent contrasts. For example, we may consider the comparison of distribution of Y in group t ($t \neq 0$) versus control group (i.e., $t = 0$). That is, we construct $(\mathcal{M} - 1)$ U-statistics

$$\boldsymbol{\mathcal{U}}_a = \begin{pmatrix} \mathcal{U}_a^{(0,1)} \\ \mathcal{U}_a^{(0,2)} \\ \vdots \\ \mathcal{U}_a^{(0,\mathcal{M}-1)} \end{pmatrix} \tag{3.21}$$

One may also consider Kruskal-Wallis test in the form specified as Satten et al. (2018) as group t versus all other group for $t = 1, \ldots, \mathcal{M} - 1$ or Jonkheere-Terpstra as groups $(0, \ldots, t)$ versus $(t + 1, \ldots, \mathcal{M} - 1)$ for $t = 0, \ldots, \mathcal{M} - 2$. However, the two specifications don't provide superiority score for group comparisons. Thus, we use the $(\mathcal{M} - 1)$ U-statistics in equation (3.21) to test the overall treatment effect and provide the pairwise comparisons. Similar to equation (3.19), the $(\mathcal{M}-1)$ U-statistics

in equation (3.21) can be decomposed as

$$\boldsymbol{\mathcal{U}}_a = \begin{pmatrix} \mathcal{U}_a^{(0,1)} \\ \mathcal{U}_a^{(0,2)} \\ \vdots \\ \mathcal{U}_a^{(0,\mathcal{M}-1)} \end{pmatrix} = \begin{pmatrix} \sum_{i:T_i=0} \xi_i^{(1)} + \sum_{j:T_j=1} \xi_j^{(1)} \\ \cdots \\ \sum_{i:T_i=0} \xi_i^{(\mathcal{M}-1)} + \sum_{j:T_j=\mathcal{M}-1} \xi_j^{(\mathcal{M}-1)} \end{pmatrix}$$

Satten et al. (2018) established the asymptotic normality for $\boldsymbol{\mathcal{U}}_a$. That is, $\boldsymbol{\mathcal{U}}_a$ is asymptotically normally distributed with mean $\boldsymbol{\mu}$ and variance V:

$$\sqrt{N}\left(\boldsymbol{\mathcal{U}}_a - \boldsymbol{\mu}\right) \sim MVN(0, V).$$

Under the null hypothesis that there is no group difference among $\mathcal{M}$ groups, each component of $\boldsymbol{\mathcal{U}}_a$ has a mean $\frac{1}{2}$ and variance as defined in (3.20). The covariance between $\mathcal{U}_a^{(0,t)}$ and $\mathcal{U}_a^{(0,t')}$, say

$$v_{tt'} = Cov\left(\sum_{i:T_i=0} \xi_i^{(t)} + \sum_{j:T_j=t} \xi_j^{(t)}, \sum_{i':T_{i'}=0} \xi_{i'}^{(t')} + \sum_{j':T_{j'}=t'} \xi_{j'}^{(t')},\right)$$

may be approximated by

$$\begin{aligned} v_{tt'} = &\sum_{i:T_i=0} \sum_{i':T_{i'}=0} \left(\xi_i^{(t)} - \overline{\xi}_0^{(t)}\right)\left(\xi_{i'}^{(t')} - \overline{\xi}_0^{(t')}\right) + \sum_{i:T_i=0} \sum_{j':T_{j'}=0} \left(\xi_i^{(t)} - \overline{\xi}_0^{(t)}\right)\left(\xi_{j'}^{(t')} - \overline{\xi}^{(t')}\right) \\ &+ \sum_{j:T_j=0} \sum_{i':T_{i'}=0} \left(\xi_j^{(t)} - \overline{\xi}^{(t)}\right)\left(\xi_{i'}^{(t')} - \overline{\xi}_0^{(t')}\right) + \sum_{j:T_j=0} \sum_{j':T_{j'}=t'} \left(\xi_j^{(t)} - \overline{\xi}^{(t)}\right)\left(\xi_{j'}^{(t')} - \overline{\xi}^{(t')}\right). \end{aligned} \tag{3.22}$$

Wald test (Test statistic: $W = \left(\boldsymbol{\mathcal{U}}_a - \boldsymbol{\mu}\right)' V^{-1}\left(\boldsymbol{\mathcal{U}}_a - \boldsymbol{\mu}\right)$) can be used to test whether there are treatment difference among the $\mathcal{M}$ groups. However, we found the test statistic using the covariance (3.22) leads to larger actual size than the nominal size

0.05 under the null hypothesize. We propose using bootstrap sampling to obtain the variance matrix, then apply the Wald test. The bootstrap method for estimating variance matrix can be carried out by the following steps:

Step 1: Draw N indices from $\{1, 2, \ldots, N\}$ with replacement and obtain the bootstrap sample by selecting those indexed rows from the original data.

Step 2: Calculate the adjusted U-statistic in (3.21) using the bootstrap sampled.

Step 3: Repeat Step 1 and Step 2, say 200 times, and obtain 200 adjusted U-statistics.

Step 4: Calculate the empirical variance of the 200 estimated U-statistics, which is considered as the variance estimate of the U-statistic $\boldsymbol{\mathcal{U}}_a$.

3.3.3 GPS-based regression

GPS-based regression and stratification have been popular for estimating ATE for two groups when outcome is continuous. Yang et al. (2016) extended GPS-based stratification to multiple groups for continuous outcomes, which can not be directly applied to estimate superiority score for ordinal outcome. We propose GPS-based regression and stratification approach to estimate ATE when outcome is ordinal. We assume that the GPS is estimated by the ordinal logistic regression model in (3.11). We take $X'\hat{\boldsymbol{\beta}}$ as the basic building block for regression and stratification. Without loss of generality, let denote $\hat{e} = X'\hat{\boldsymbol{\beta}}$. The GPS-based regression uses GPS, say $\hat{e}$, instead of the covariates in the parametric models specified in Section 3.2.2. GPS has been referred as a dimension reduction tool (Guo, 2010). In GPS-based regression methods, we fit a cumulative link model as given in equation (3.9) but with control of $\hat{e}$ instead of the covariates:

$$\mathcal{F}^{-1}(P[Y \leq j|X, T]) = \alpha_j - \tau_1 T^{(2)} - \ldots - \tau_{\mathcal{M}-1} T^{(\mathcal{M})} - \delta\hat{e}, \quad j = 1, \ldots, \mathcal{C}-1. \quad (3.23)$$

Estimation of superiority score and testing of hypothesis of treatment effect follow the same procedure as described in Section 3.2.2. When the number of covariates is large, we expect that the GPS-based regression performs better than the regular parametric model.

3.3.4 GPS-based stratification

We propose GPS-based stratification by using the GPS estimated from ordinal logit model (3.11). GPS is considered as a balance score. We divide the entire data in multiple strata, say $S = 5$, based on the quantiles of GPS (say $\hat{e}$). Within each stratum, we hope that the covariates from different treatment groups become balanced. The GPS-based stratification are carried out by the following steps:

Step 1: Estimate the GPS based on the ordinal logistic model in (3.11). Form S strata based on the sample quantiles of $\hat{e}$.

Step 2: At s^{th} stratum, estimate the superiority score (say, $\hat{\gamma}_{tt'}^{(s)}$) between t^{th} and t'^{th} treatment groups where $t \neq t'$ and $t, t' \in \{0, \ldots, \mathcal{M}-1\}$.

Step 3: Calculate the variance estimate for the superiority score $\hat{\gamma}_{tt'}^{(s)}$ at s^{th} stratum, denoted as $\widehat{var}(\hat{\gamma}_{tt'}^{(s)})$, which are obtained from the Hoeffding projection.

Step 4: Calculate the overall estimate of the superiority score between t^{th} and t'^{th} treatment as the weighted mean of the superiority scores over S strata: $\hat{\gamma}_{tt'} = \sum_{s=1}^{S}(N_s/N)\hat{\gamma}_{tt'}^{(s)}$, where N_s is the sample size of s^{th} stratum, and N is the total sample size. The variance for $\hat{\gamma}_{tt'}$ is estimated as $\widehat{var}(\hat{\gamma}_{tt'}) = \sum_{s=1}^{S}(N_s/N)^2\widehat{var}(\hat{\gamma}_{tt'}^{(s)})$.

In Step 2, the superiority scores can be estimated parametrically or nonparametrically using U-statistics. In the simulation and case study, we have used stratified U-statistic to estimate the superiority score. The superiority score between t^{th} and

t'^{th} treatment within the s^{th} stratum based on stratified U-statistic is obtained as

$$\hat{\gamma}_{tt'}^{(s)} = \frac{1}{\sum_i I(T_i = t, S_i = s)} \frac{1}{\sum_j I(T_j = t', S_j = s)} \sum_{i \in \{i: T_i = t, S_i = s\}} \sum_{j \in \{j: T_j = t', S_j = s\}} \big[P(Y_i < Y_j) + \frac{1}{2} P(Y_i = Y_j)\big],$$

where $t, t' \in \{0, \ldots, \mathcal{M} - 1\}$, and $t \neq t'$.

3.3.5 Covariate balances

In practice, it is essential to check whether the GPS balances the covariates among different treatment groups. The popular metric for checking the balance of covariates is the absolute standardized mean differences (ASMD) (McCaffrey et al., 2013). The population standardized bias (PSB) for p^{th} covariate and t^{th} treatment is calculated as

$$PSB_{p,t} = \left\{ \frac{|\overline{X}_p^{(t)} - \overline{X}_p|}{\hat{\sigma}_p} \right\},$$

where $\overline{X}_p$ and $\hat{\sigma}_p$, respectively, denote the unweighted mean and standard deviation of the p^{th} covariate for the pooled sample across all treatments, and $\overline{X}_p^{(t,w)} = \frac{\sum_{i=1}^N w_i X_{i,p} I\{T_i = t\}}{\sum_{i=1}^N w_i I\{T_i = t\}}$, is the weighted average of p^{th} covariate for the subjects from t^{th} treatment group using weights w_i $(i = 1, \ldots, N)$. The weights are defined as the inverse of the probability of the treatment the subject received. The overall balance for the p^{th} covariate is characterized by ASMD:

$$ASMD_p = \sup_{t \in \{0,1,\ldots,\mathcal{M}-1\}} \{PSB_{p,t}\}. \tag{3.24}$$

The p^{th} covariate is considered to be balanced if $ASMD_p \leq 0.1$. The covariate balance for the original data can be calculated using the same metric but with weight $w_i = 1$. Since we use either the ordinal logistic regression or multinomial regression

to estimate GPS, we have two choices of w_i. The balance of the covariates resulted from different GPS estimation methods can provide information on which GPS-based method provides a better balance of covariates.

3.4 Simulation studies

To examine the performance of the proposed methods as well as existing methods for estimating ATE for ordinal outcome, we constructed multiple simulation settings under treatment groups $\mathcal{M} = 4$ and response categories $\mathcal{C} = 4$. The treatment assignment is generated from one of the two models:

GPS1: Ordinal logistic regression model

$$\log \frac{P[T \leq t|\mathbb{X}]}{P[T > t|\mathbb{X}]} = \alpha_t + \mathbb{X}'\beta, \ t = 0, 1, 2. \tag{3.25}$$

GPS2: Multinomial regression model

$$\log \frac{P[T = t|\mathbb{X}]}{P[T = 0|\mathbb{X}]} = \alpha_t + \mathbb{X}'\beta_t, \ t = 1, 2, 3. \tag{3.26}$$

With the two GPS models, we are able to examine how the estimated GPS impacts the accuracy of the ATE estimates. We also generated the ordinal outcome from one of the following three models:

OR1: The outcome was generated from the ordinal logistic regression model:

$$\log \left\{ \frac{P[Y(t) \leq k|\mathbb{X}, T]}{P[Y(t) > k|\mathbb{X}, T]} \right\} = \alpha_k - \tau_1 T^{(1)} - \tau_2 T^{(2)} - \tau_3 T^{(3)} - \mathbb{X}'\delta, \ \text{for } k = 1, 2, 3. \tag{3.27}$$

Here $(\alpha_1, \alpha_2, \alpha_3)$ are chosen as $(-\log(3), 0, \log(3))$ so that the probability of occurrence of each level of response is same.

OR2: The outcome was generated from the probit regression model:

$$\Phi^{-1}\left(P[Y(t) \leq k|\mathbb{X}, T]\right) = \alpha_k - \tau_1 T^{(1)} - \tau_2 T^{(2)} - \tau_3 T^{(3)} - \mathbb{X}'\delta, \text{ for } k = 1, 2, 3. \tag{3.28}$$

OR3: The outcome was generated from a mixture distribution of the Box-cox family of the following form

$$Y_i = \sum_{t=1}^{\mathcal{M}} F_t(\mathbb{X}_i, T_i) I\{T_i = t\}, \tag{3.29}$$

where $F_t(\mathbb{X}_i, T_i) = F(\alpha_k + \tau_t + \mathbb{X}\delta; \lambda_t)$, and F is the CDF of the Box-cox family (Guerrero and Johnson, 1982) of the following form

$$F(x;\lambda) = \begin{cases} 0, & \text{if } x < -\dfrac{1}{\lambda},\ \lambda > 0; \\ \dfrac{(1+\lambda x)^{\frac{1}{\lambda}}}{1+\lambda x)^{\frac{1}{\lambda}} + 1}, & \text{if } 1 + \lambda x > 0,\ \lambda \neq 0; \\ 1, & \text{if } x > -\dfrac{1}{\lambda},\ \lambda < 0. \end{cases} \tag{3.30}$$

Here $\lambda_t = 1 + \tau_t$, and $(\alpha_1, \alpha_2, \alpha_3)$ is set such that the probabilities of occurrence at four levels of responses as (02,0.2,0.2,0.4).

3.4.1 Simulation scenarios

In this section, we carry out extensive simulations to examine the performance of different methods when treatment is generated from one of the two models (GPS1 and GPS2), and outcome is generated from one of the three models (OR1, OR2, and OR3). Under each combination of GPS-model and outcome model, we generated 1000 simulated data. For each simulated data, we carried out hypothesis test and estimated ATE (i.e., superiority score). The simulation procedure is described below:

Step 1: Generate a sample of size N (say $N = 5000$) for the covariates, say $\mathbb{X} =$

$(X_1, X_2, \ldots, X_{12})$, where $(X_1, \ldots, X_6) \sim MVN(\underline{0}, \Sigma_1)$ and $(X_7, \ldots, X_{12}) \sim MVN(\underline{0}, \Sigma_2)$, where $\Sigma_1 = \begin{bmatrix} 1 & \rho_1 & \rho_1 & \ldots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \ldots & \rho_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \ldots & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & \rho_2 & \rho_2 & \ldots & \rho_2 \\ \rho_2 & 1 & \rho_2 & \ldots & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_2 & \rho_2 & \rho_2 & \ldots & 1 \end{bmatrix}$, $\rho_1 = 0.1$, and $\rho_2 = 0.2$.

Step 2: Generate treatment assignments based on either the ordinal logistic model (GPS1) or the multinomial regression model (GPS2). In the ordinal logistic regression model, the covariate parameter $\boldsymbol{\beta}$ is set as (-0.1, -0.2, -0.3, -0.3, -0.2, -0.1, 0, 0, 0). In the multinomial regression model we set the vectors of parameters for $\boldsymbol{\beta}$ as

$$\begin{pmatrix} \beta_2' \\ \beta_3' \\ \beta_4' \end{pmatrix} = \begin{bmatrix} -0.1 & -0.1 & -0.1 & 0 & 0 & 0 & -0.1 & -0.1 & -0.1 & 0 & 0 & 0 \\ -0.1 & -0.1 & -0.1 & 0 & 0 & 0 & -0.1 & -0.1 & -0.1 & 0 & 0 & 0 \\ -0.1 & -0.1 & -0.1 & 0 & 0 & 0 & -0.1 & -0.1 & -0.1 & 0 & 0 & 0 \end{bmatrix}$$

Step 3: Generate outcome for each observed $\mathbb{X}$ and T by one of the three outcome models (i.e., OR1, OR2, and OR3). In each model, we set δ =(0.1, -0.2, -0.3, -0.2, -0.3, -0.1, 0, 0, 0, 0, 0, 0) and set $(\tau_2, \tau_3, \tau_4) = (c, 2c, 3c)$. Here c reflects the magnitude of the treatment effect and c is set as a sequence from 0 to 1 with 0.05 increment.

Step 4: Construct the test statistic to test whether there is treatment effect, and estimate the superiority score for each comparison based on each method.

Step 5: Repeat Step 1 through Step 4 1000 times.

We summarize the simulation result by (1) calculating the percentage of rejection under the null hypothesis that there is no treatment effect among the $\mathcal{M}$ groups; (2) calculating the average of the estimated superiority score for each pair of treatment

groups; (3) calculating the mean of the standard error estimate for each superiority score, as well as the empirical standard deviation for each of the 1000 estimated superiority score.

3.4.2 Simulation results

We first reported the simulation results on the overall significant test when the treatment assignment was generated from ordinal logistic regression model and the outcome was generated from one of the outcome models (Table 3.1). All the GPS-based methods, parametric or adjusted U-statistics, all have the actual sizes close to the nominal size 0.05. However the size from the unadjusted U-statistic is far off from the nominal size 0.05. We also plotted the power curves of all different methods in Figure 3.1. The parametric methods seem to have higher power than the adjusted U-statistics for testing the treatment effects.

However, when comparing the superiority scores, the estimates from the GPS based parametric methods have large biases. Figures 3.2, 3.4, and 3.6 show the biases from the true superiority scores under different models, when the response are generated from Boxcox (Figure 3.2), logit (Figure 3.4) and probit models (Figure 3.6) respectively. It is clear that the bias under parametric models increases as the treatment effect increases. On the other hand, the mean square errors (Figures 3.3, 3.5, 3.7) using the GPS adjusted U-statistic methods are the lowest. The regular nonparametric method has large biases for estimating the superiority scores.

We also ran simulations when the treatment was generated from multinomial regression (i.e., equation (3.26)) and for highly correlated covariates. The simulation results are provided in Appendix A.4, which are similar to what we got for weakly correlated covariates. From the simulation results presented in this section and in Appendix A.4, we conclude that the estimation of superiority scores from ordinal cumulative regression or GPS-based ordinal cumulative regression are biased no matter

how powerful the associated test statistics are. Such bias may be due to the strong assumption on variance by Agresti and Kateri (2017). Based on our simulation study, when there is no treatment effect, the parametric models successfully estimates the superiority scores. However, as the treatment effect increased, the parametric models led to increased bias regardless what the treatment and response generating models were. On the other hand, the adjusted U-statistics resulted in nearly unbiased estimates. The biases incurred using stratified U-statistic are much smaller than those from the parametric models. However, the estimates from the stratified U-statistic are not as good as the adjusted U-statistic models. We conclude that the regular and GPS based parametric models are powerful tools to test the treatment effect. However, to quantify the treatment effect, the adjusted U-statistics are better choices for ordinal outcome.

3.5 Case Study

In this section we apply the methods described in Section 3.3 to the National Health and Nutrition Examination Survey (NHANES) 2011-2014 data. The goal of this case study is to investigate whether toxic metals such as cadmium and inorganic arsenic are potential risk factors for the chronic kidney disease (CKD). CKD is a condition characterized by a gradual loss of kidney function over time, which is defined based on the urine albumin to creatinine ratio (ACR) and glomerular filtration rate (GFR) (Levey et al., 2002). GFR is estimated based on serum creatinine, age, race, and sex of the subject using the CKD Epidemiology (CKD-EPI) equation (Levey et al., 2009):

$$eGFR = 141 \times min(S_{cr}/k, 1)^{\alpha} \times max(S_{cr}/k, 1)^{-1.209} \times 0.993^{Age} \times 1.018\,[\text{if female}] \times 1.018\,[\text{if African American}]\,, \tag{3.31}$$

where S_{cr} is serum creatinine in mg/dL, k is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males. High albumin in urine (i.e., large ACR) is an early sign of kidney damage, and low $eGFR$ indicates impaired kidney function. Based on different values of ACR and $eGFR$, the prognosis of CKD can be classified as four different levels: very high-risk CKD, high-risk CKD, moderate-risk CKD, and no CKD (see Figure 3.8).

Cadmium and arsenic are toxic metals that have wide distributions in the environment, known to damage multiple organs even at lower levels of exposure. Cadmium can cause significant damage to kidneys (Adams et al., 1969; Garçon et al., 2007). Inorganic arsenic is known to be associated with renal injury (Giberson et al., 1976; Ratnaike, 2003; Vaziri et al., 1980). The co-exposure of cadmium and arsenic is believed to cause more pronounced renal toxicity than exposure to each of the agents alone. We hypothesize that high level of arsenic and cadmium exposures are associated with high prognosis of CKD. We consider four exposure groups: low arsenic and low cadmium (low-low), low arsenic and high cadmium (low-high), high arsenic and low cadmium (high-low), and high arsenic and high cadmium (high-high). We categorize exposure of each metal as low if the amount of it in each subject is less than the 67^{th} percentile of the entire sample, otherwise we categorize it as high exposure.

To examine the association between metal exposure and CKD, we control the demographic information: sex, age, race, marital status, education level, hypertension status. We obtained 3181 subjects who had complete information on demographics and metal exposure. Observed frequencies based on toxic metal exposure and risk of CKD are presented in Table 3.2. From Table 3.2, regardless the level of arsenic, the percentages of people with no risk of CKD are lower when cadmium levels are higher. It seems that cadmium level is associated with risk of CKD. We performed the χ^2-test, which indicates that the risk of CKD is significantly associated with arsenic-cadmium exposure ($p < 0.01$). In particular, high level of cadmium exposure is associated with

higher rate of risk of CKD.

We also summarized the demographic information across different exposure groups in Table 3.3. From Table 3.3, we observe that high cadmium exposure group tends to include older people. We also observe that almost all demographic variables are significantly different among the four exposure groups. The ordinal logistic model by regressing CDK risk status on metal exposure and demographic variables indicate that age, hypertension status, and education level are significant for CKD status. Age is found to be significantly associated with CKD (OR=0.956, p-value < 0.001), hypertension is also found to be detrimental towards the worse outcome of CKD (OR=0.601, p-value < 0.001). Education level is found to be protective to the risk of CKD (OR=1.655, p-value < 0.001).

The observed association between risk of toxic metal exposures and CKD (Table 3.2) maybe misleading. We applied the proposed GPS-based methods to examine the association between toxic metal exposure and risk of CKD with control of confounding variables. Table 3.4 summarizes the test statistic for whether there are group differences, also the estimates of the superiority scores between two groups and the associated p-values on whether the superiority scores are 0.5. From Table 3.4, based on different methods, the conclusions for the association between toxic metal exposure and the risk of CKD are different. One way to examine which method is more suitable is to examine the balance of confounding covariates.

We first calculated the average standardized mean difference (ASMD) for the original observed data (see Figure 3.9). Then we calculated ASMD using the weights obtained from multinomial model and ordinal logistic model. From Figure 3.9, the GPS calculated from ordinal logistic model failed to balance the covariates properly. On the other hand, the GPS from the multinomial regression have balanced the covariates (i.e. ASMD < 0.1 for all confounding variables). The parametric models are unable to compute the superiority score correctly. Therefore, in this case study,

we draw conclusion based on adjusted U-statistics where the GPS was obtained from the multinomial regression. From Table 3.4 we conclude that we did not find any significant association between cadmium-arsenic exposure and risk of CKD.

3.6 Conclusion and discussion

This study unifies popular parametric and nonparametric GPS-based approaches to estimate ATE when response is ordinal. To assess the treatment effects of multiple groups we propose using superiority scores. The simulation results under different scenarios exhibited the supremacy of the adjusted U-statistics. The adjusted U-statistics showed better performance in estimating the superiority scores with lower mean squared error. The parametric models, despite how powerful they are, incurred large bias due to strong variance assumption. The case study is a perfect example to portray how important the covariate balance is. Since the true nature of treatment effect is unknown, it is essential to draw conclusion based on the right model chosen by the covariate balance. The case study didn't show any significant association of cadmium and arsenic exposure on the prognosis of CKD status.

3.7 Tables and Figures

Table 3.1: Rejection rate for the overall test for different methods, where the treatment assignment was generated from OLR, and outcome variable was generated from ordinal logit model (OR1), probit model (OR2), and Box-cox model (OR3), respectively

Methods	Outcome models		
	Logit	**Probit**	**Boxcox**
GPS regression (logit)	0.054	0.044	0.043
GPS regression (probit)	0.055	0.042	0.043
Parametric regression (logit)	0.049	0.055	0.051
Parametric regression (probit)	0.051	0.055	0.049
Adjusted U (Multinomial)	0.053	0.056	0.053
Adjusted U (OLR)	0.060	0.051	0.056
Unadjusted U (KW)	0.783	0.996	0.864

Table 3.2: Observed frequency based on toxic metal exposure and the risk of CKD

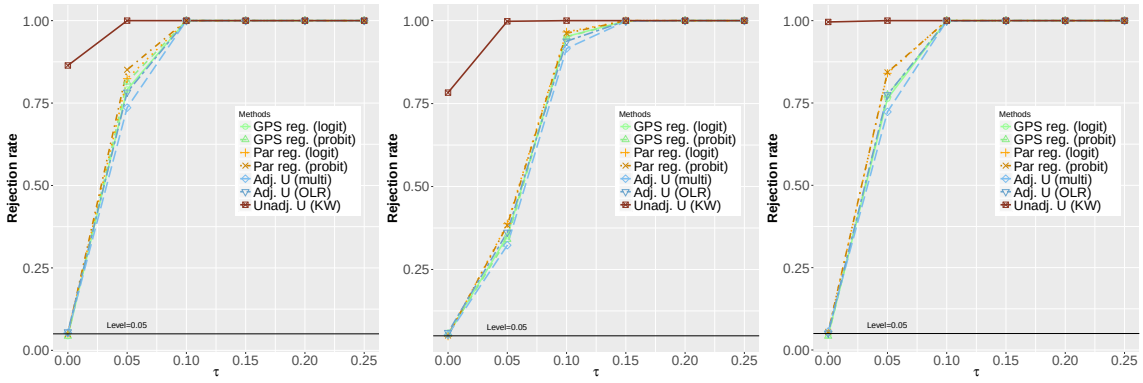
Arsenic & Cadmium	Response Levels (i.e. risk of CKD)			
	Very high-risk	**High-risk**	**Moderate-risk**	**No CKD**
Low-Low	26 (1.97%)	29 (2.2%)	108 (8.18%)	1158 (87.66%)
Low-High	26 (3.33%)	31 (3.97%)	125 (16.03%)	598 (76.67%)
High-Low	11 (2.58%)	6 (1.41%)	27 (6.34%)	382 (89.67%)
High-High	15 (2.29%)	21 (3.21%)	85 (13%)	533 (81.5%)

Table 3.3: The summarized demographic variables across different cadmium and arsenic exposure groups

Variables	Treatment groups				
	Low-Low N=1321	**Low-High** N=780	**High-Low** N=426	**High-High** N=654	**p-value**
Continuous variables	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	**Mean (SD)**	
Age	43.87 (16.97)	56.16 (15.26)	39.97 (16.18)	52.89 (16.56)	<0.001
Categorical Variables	**n (%)**	**n (%)**	**n (%)**	**n (%)**	
Sex = Male	636 (48.1%)	350 (44.9%)	263 (61.7%)	340 (52.0%)	<0.001
Race= White	639 (48.4%)	359 (46.0%)	129 (30.3%)	177 (27.1%)	<0.001
Race= Black	229 (17.3%)	212 (27.2%)	81 (19.0%)	176 (26.9%)	<0.001
Hispanic = Yes	296 (22.4%)	143 (18.3%)	114 (26.8%)	136 (20.8%)	0.006
Education Level= High	1087 (82.3%)	559 (71.7%)	350 (82.2%)	488 (74.6%)	<0.001
Married=Yes	640 (48.4%)	391 (50.1%)	212 (49.8%)	348 (53.2%)	0.263
Hypertension = Yes	444 (33.6%)	346 (44.4%)	125 (29.3%)	248 (37.9%)	<0.001

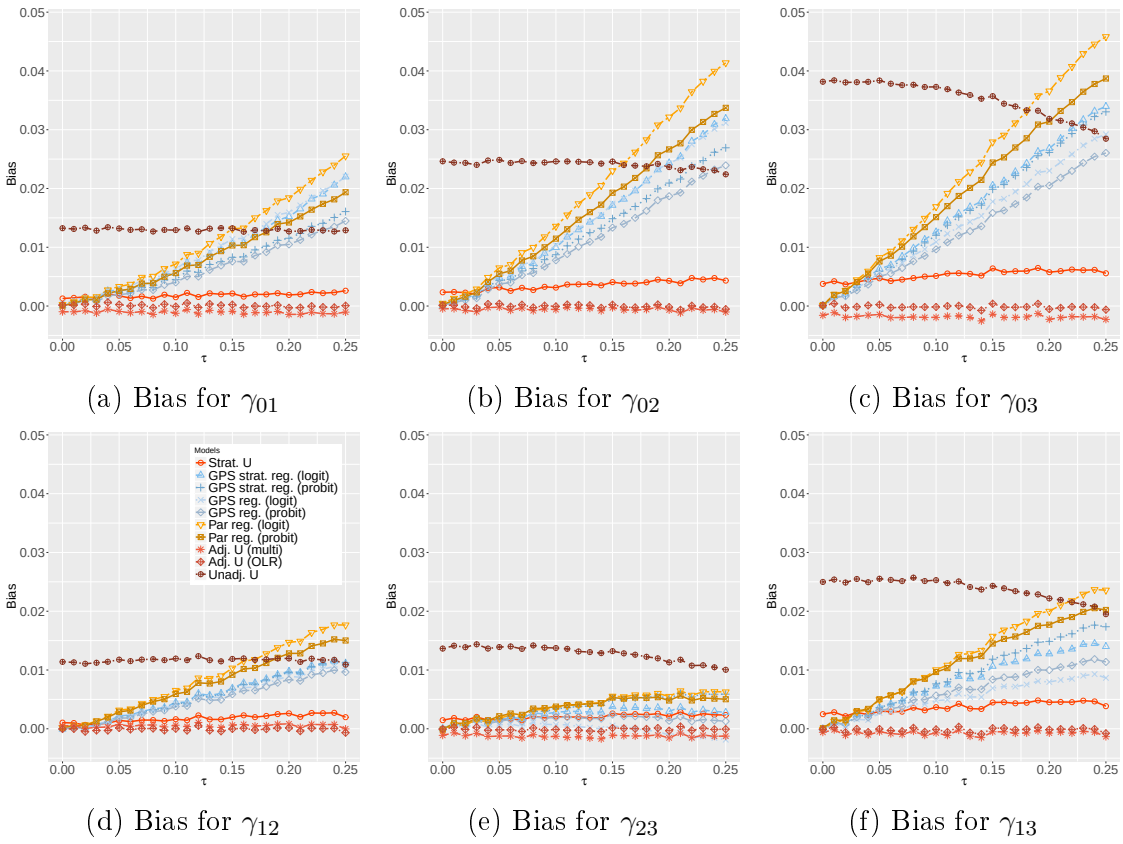
Table 3.4: Superiority score estimates

	Unadj. U		Adj. U (Multi.)		Adj. U (OLR)		Strat. U		Strat. GPS Reg. (Logit)		GPS reg. (Logit)	
Overall	Test Stat.	p	Test Stat.	p	Test Stat.	p	Test Stat.	p	Test Stat.	p	Test Stat.	p
test	55.82	<0.001	1.79	0.616	38.40	<0.001	-	-	-	-	34.63	<0.001
Sup. score	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p
γ_{01}	0.45	0.040	0.49	0.214	0.45	<0.001	0.46	0.170	0.42	0.117	0.40	0.016
γ_{02}	0.51	0.738	0.49	0.616	0.52	0.027	0.52	0.602	0.55	0.467	0.55	0.404
γ_{03}	0.47	0.260	0.50	0.969	0.49	0.289	0.49	0.667	0.47	0.603	0.48	0.661
γ_{12}	0.56	0.024	0.50	0.798	0.57	<0.001	0.55	0.082	0.62	<0.001	0.65	<0.001
γ_{23}	0.46	0.260	0.51	0.646	0.47	0.004	0.47	0.445	0.43	0.083	0.43	<0.001
γ_{13}	0.52	0.370	0.51	0.267	0.54	0.001	0.53	0.394	0.55	0.102	0.58	<0.001



(a) Response from Boxcox model (b) Response from Logit model (c) Response from Probit model

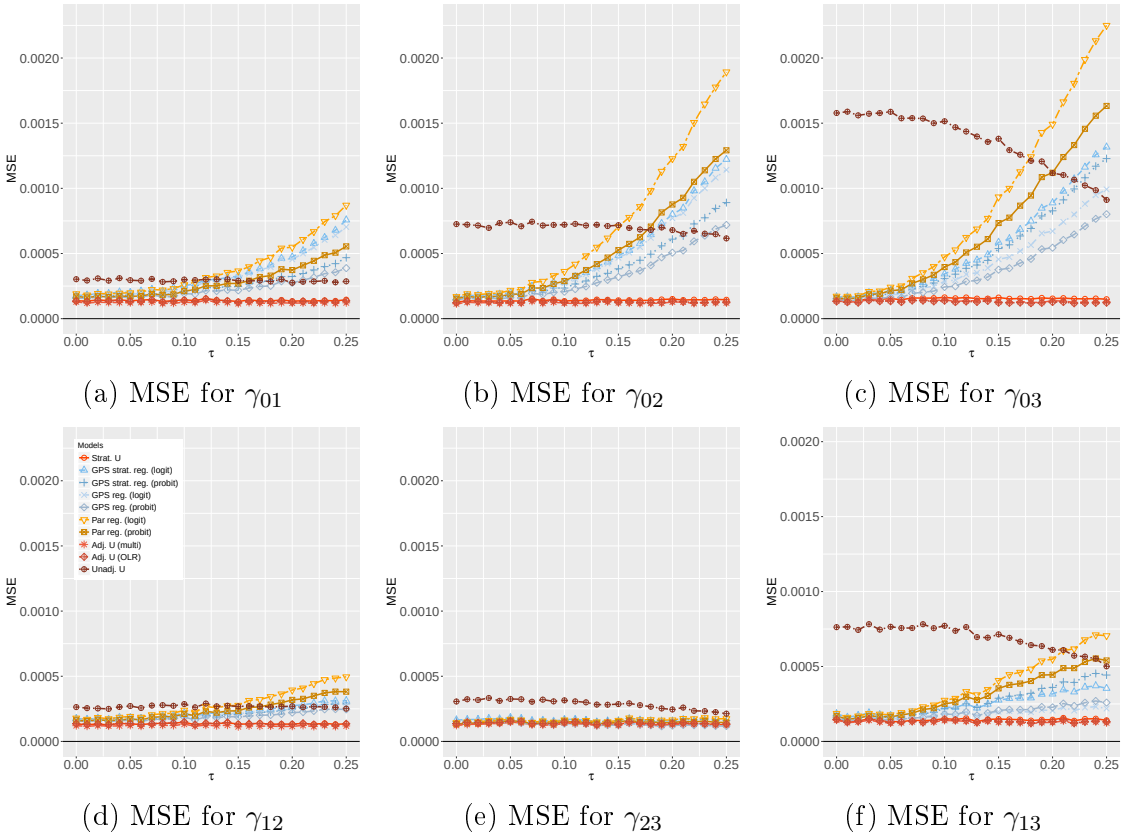
Figure 3.1: Power curve for all methods when treatment was generated from OLR



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

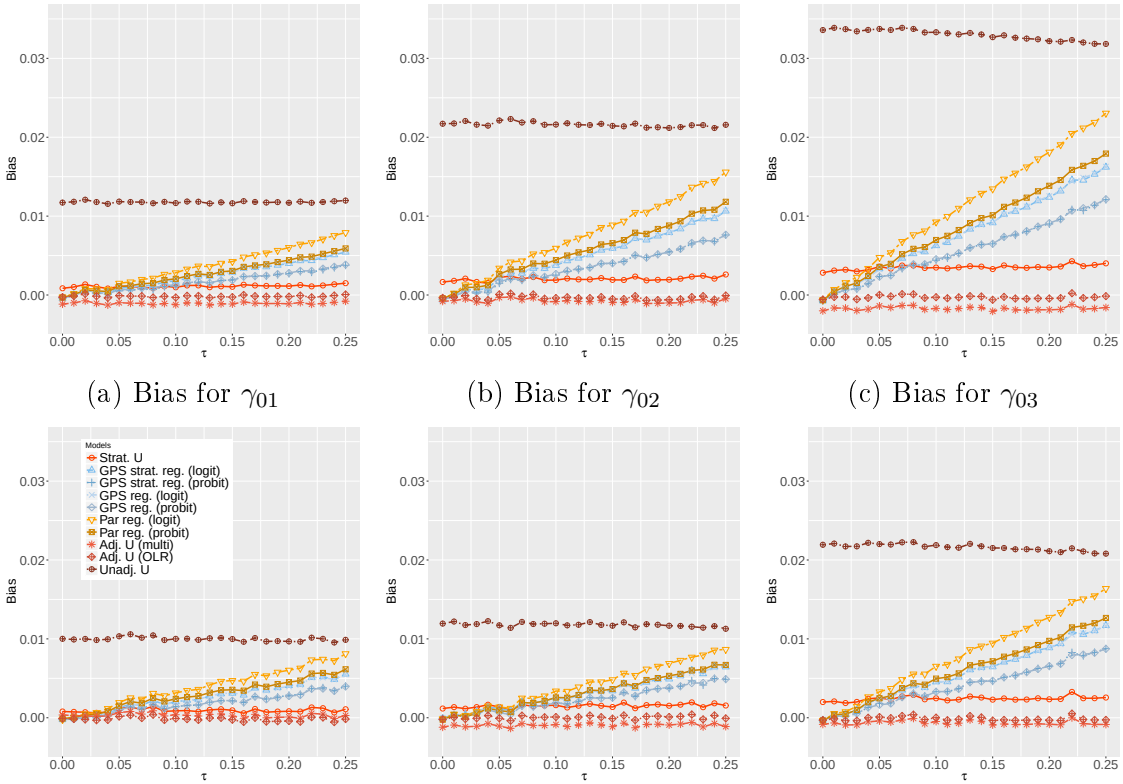
Figure 3.2: Bias plot for superiority scores when outcome was from Boxcox model and treatment was generated from ordinal logit model



(a) MSE for γ_{01} (b) MSE for γ_{02} (c) MSE for γ_{03}

(d) MSE for γ_{12} (e) MSE for γ_{23} (f) MSE for γ_{13}

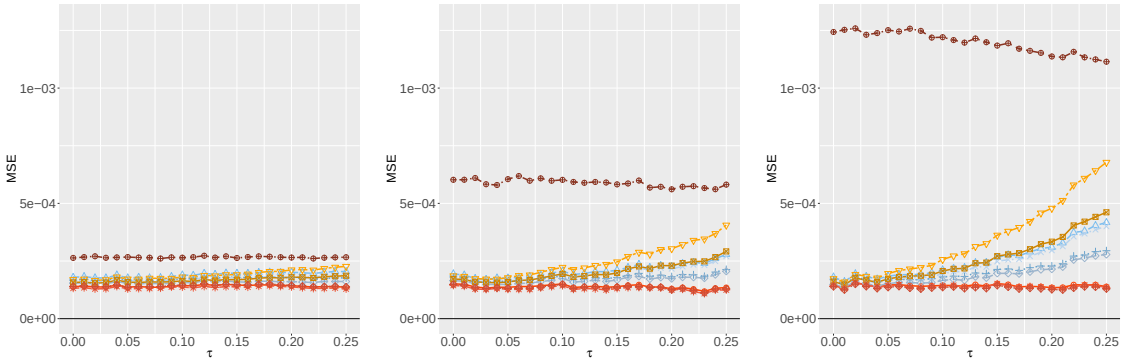
Figure 3.3: MSE plot for superiority scores when outcome was from Boxcox model and treatment was generated from ordinal logit model



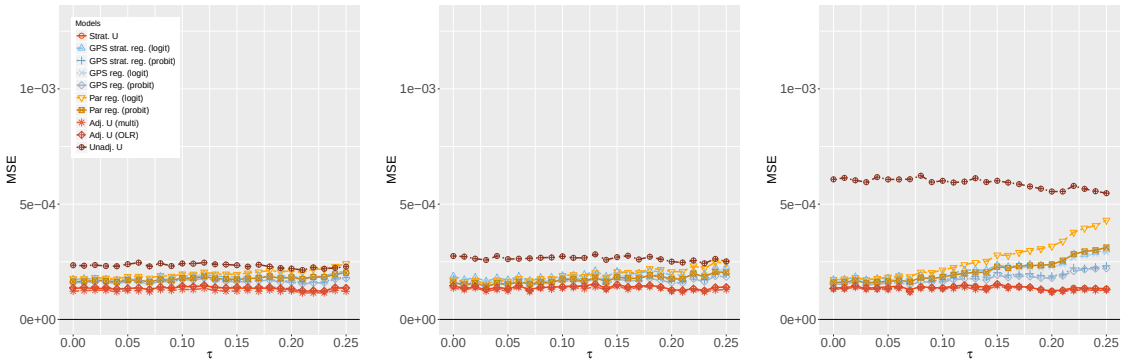
(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

Figure 3.4: Bias plot for superiority scores when outcome was from ordinal logit model and treatment was from ordinal logit model

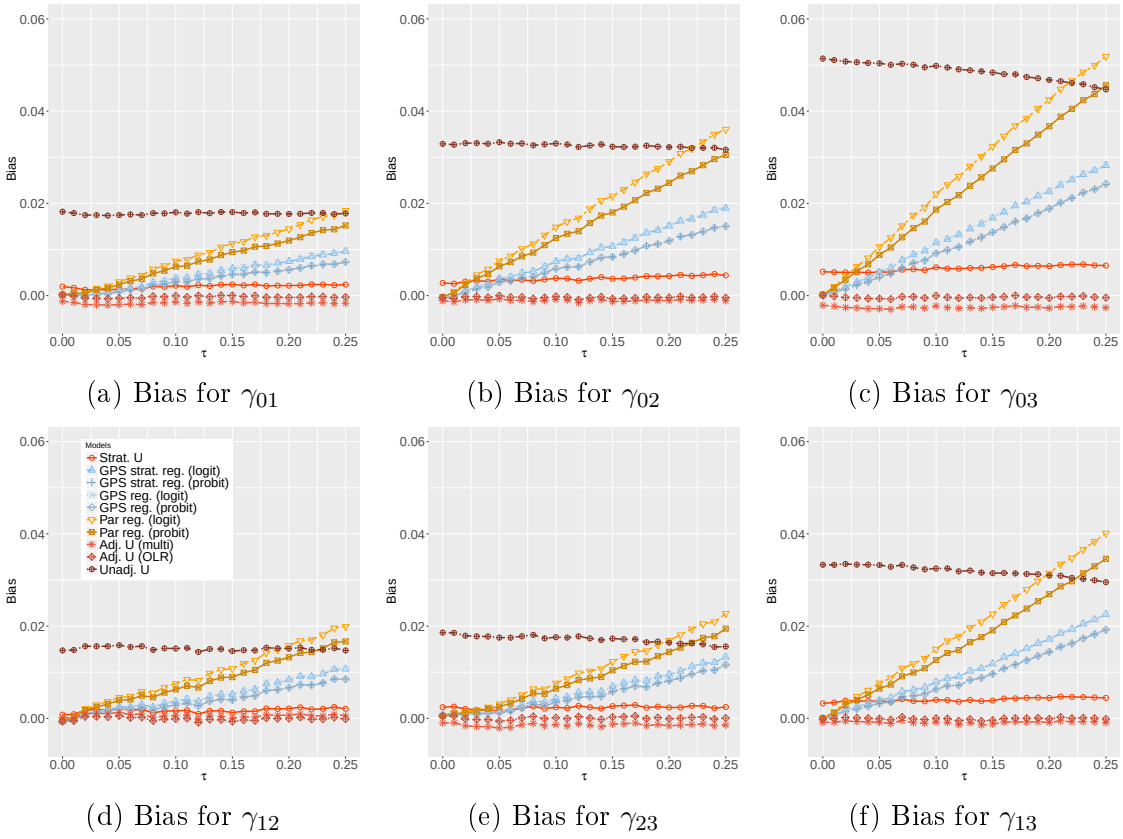


(a) MSE for γ_{01} (b) MSE for γ_{02} (c) MSE for γ_{03}



(d) MSE for γ_{12} (e) MSE for γ_{23} (f) MSE for γ_{13}

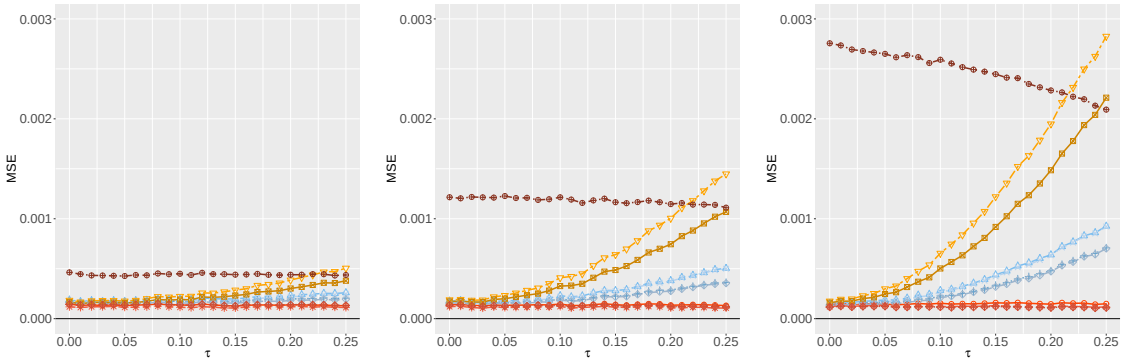
Figure 3.5: MSE plot for superiority scores when outcome was from ordinal logit model and treatment was generated from ordinal logit model



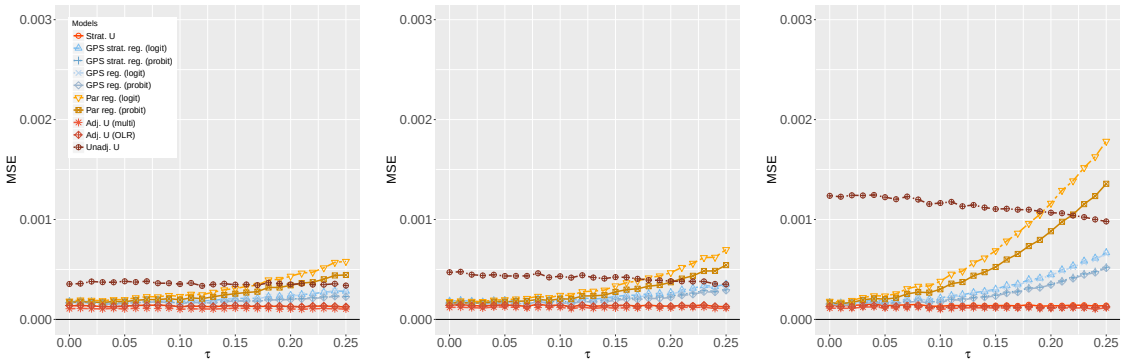
(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

Figure 3.6: Bias plot for superiority scores when outcome was from ordinal probit model and treatment was generated from ordinal logit model



(a) MSE for γ_{01} (b) MSE for γ_{02} (c) MSE for γ_{03}



(d) MSE for γ_{12} (e) MSE for γ_{23} (f) MSE for γ_{13}

Figure 3.7: MSE plot for superiority scores when outcome was from probit model and treatment was generated from ordinal logit model

No CKD
Moderate-risk CKD
High-risk CKD
Very high-risk CKD

GFR stages, description, and range (mL/min per 1.73m^2)				Albuminuria stages, description, and range (mg/g)				
				A1		A2	A3	
				Optimum and high-normal		High	Very high and nephrotic	
				< 10	10-29	30-299	300-1999	≥ 2000
	G1	High and optimum	< 105					
			90-104					
	G2	Mild	75-89					
			60-74					
	G3a	Mild-moderate	45-59					
	G3b	Mild-severe	30-44					
	G4	Severe	15-29					
	G5	Kidney failure	< 15					

Figure 3.8: CKD prognostic status based on ACR and $eGFR$ (Image taken from Levey and Coresh (2012))

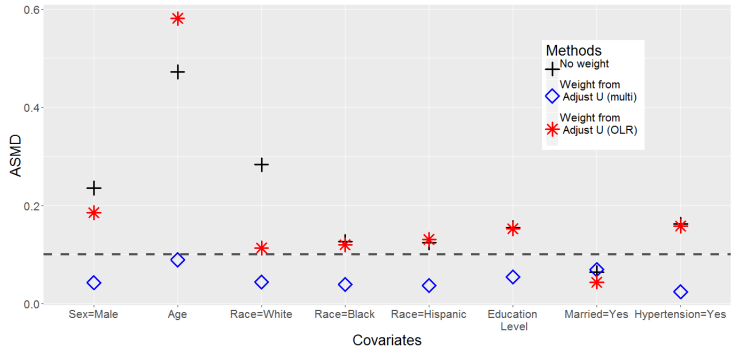


Figure 3.9: Balance of covariates using weights from multinomial and ordinal logistic regression

CHAPTER 4

GENERALIZED SPATIOTEMPORAL ADDITIVE MODEL IMPLEMENTED IN R AND ITS APPLICATION TO ASSESSING OVERUSE OF ANTIBIOTICS DRUGS FOR UPPER RESPIRATORY TRACT INFECTIONS IN KENTUCKY

4.1 Introduction

Spatiotemporal data are data collected over time across space. Spatiotemporal data are very common in various fields such as climate science, social sciences, neuroscience, epidemiology and public health (Kang et al., 2011; Meliker et al., 2005; Peuquet and Duan, 1995). Analyzing spatiotemporal data is more challenging than analyzing purely spatial or time series data. However spatiotemporal data have the advantages that allow us to simultaneously study the time effect and geographic variation, if analysis and modeling are done appropriately.

Generalized linear model (GLM) is commonly used to model the relationship between the response and the covariates when the response follows exponential family of distributions (McCullagh, 1989; Nelder and Baker, 1972; Nelder and Wedderburn, 1972). In GLM, a parametric mean function of the response is used to link with the linear combination of the covariates. However, if the relationship between the covariates and the response variable is complicated, the GLM approach becomes inadequate (Fang and Chan, 2014). The generalized additive models (GAM) (Hastie and Tibshirani, 1986) extends the GLM by introducing more flexible nonparametric functionals

of the covariates, thus GAM can be used to model complex relationship between the response variable and covariates. In this study, we introduce Fourier functions are introduced to the GAM (Kong et al., 2012) to model time trend and seasonal variation, and we introduce thin-plate splines to model geographical variations.

The statistical model and analysis developed here are to investigate the antibiotic overuse. Antibiotics are antimicrobial drugs which are targeted to prevent the growth of bacteria. Antibiotics can neither prevent the growth of viruses (that cause common cold or influenza) nor reduce the course of acute illness caused by the viruses (Alumran et al., 2012; Soyka et al., 1975). However. the use of antibiotics as antiviral agents still remains high over the past decade, which not only increases medical cost but also increases the risk of life threatening adverse effects and microbial resistance (Gonzales et al., 1997; Zhang et al., 2012). The adverse side-effects of antibiotic overuse include gastrointestinal effects, drowsiness, diarrhea, and hyperactivity (Alumran et al., 2012). Studies have shown that upper respiratory tract infection (URI) is one of the virus borne acute infections, which don't need to be treated with antibiotics (Snow et al., 2001; Watson et al., 1999). URI is a viral syndrome of fever, rhinorrhea, cough and usually self-limiting in nature (Alumran et al., 2012; Soyka et al., 1975). Hence, prescribing antibiotics to the patients diagnosed with URI has no beneficial effect, rather harmful. Yet physicians around the world continue to prescribe antibiotics to treat viral URI (Mainous III and Hueston, 1998). Studies have showed that young children aged between 5-11 years are more prone to get antibiotic prescription for URI (Hicks et al., 2015; Nyquist et al., 1998). In 2015-2016, Kentucky had the highest overall antibiotics prescription rate (1281 prescriptions per 1000 persons) (Hicks et al., 2015).

To reduce the antibiotic misuse, it is important to identify the regions and risk factors for the misuse. In this project we develop a GAM which can not only capture the time trend and seasonal variation, but also model the geographic variation. We

present the proposed model along with the R-code in Section 4.2. We applied the proposed model to investigate antibiotic overuse in Section 4.3. The last section is devoted to a discussion.

4.2 Generalized spatiotemporal additive model

In this section, we present the proposed GAM to capture the time effect and geographic variation. We consider the outcome variable as binary outcome (e.g., antibiotic overuse or not). The approach proposed here can be applied to count data (assuming Poisson distribution) and continuous data (assuming Gaussian distribution) which can be implemented with different link function using `gam` or `bam` functions from the `mgcv` package.

Let assume that there are N medical visits from 2014 to 2016, with URI diagnosis but not having other concurrent medical conditions and competing diseases. Let Y_i denotes whether there is associated antibiotic prescription for the i^{th} medical visit. That is,

$$Y_i = \begin{cases} 1, & \text{if there is antibiotic prescription for } i^{\text{th}} \text{ visit;} \\ 0, & \text{otherwise} \end{cases}$$

Here $i = 1, \ldots, N$. The goal of this study is to examine whether the antibiotic overuse has time trend and whether the antibiotic overuse is associated with geographic variation, and demographic and socio-economic variables.

Let denote $X_i^{(t)}$ as the time elapsed in month since the start date (e.g., $X_i^{(t)}$=1 if the i^{th} visit is in January 2014; $X_i^{(t)}$=2 if the i^{th} visit is in February 2014, and etc.), $X_i^{(\text{loc})}$ as the location variables (i.e., $X_i^{(\text{loc})} = \left(X_{i,\text{lat}}^{(\text{loc})}, X_{i,\text{lon}}^{(\text{loc})}\right)$) which describes the latitude and longitudinal of the location the patient resides. Let also denote $X_i^{(D)} = \left(X_{i1}^{(D)}, \ldots, X_{ip}^{(D)}\right)$, as categorical demographic information, and $X_i^{(C)} =$

$\left(X_{i1}^{(C)}, \ldots, X_{ig}^{(C)}\right)$, as continuous demographic information and the socio-economic variables. Then the proposed GAM has the following form:

$$\log\left(\frac{P[Y_i=1|X_i]}{1-P[Y_i=1|X_i]}\right) = \beta_0 + X_i^{(D)}\beta + \sum_{j=1}^{g} f_j(x_{ij}^{(C)}) + f(X_i^{\text{loc}}) + \beta_t \frac{X_i^{(t)}}{12} + \sum_{k=1}^{6}\left[\beta_{ck}\cos\left(\frac{2k\pi X_i^{(t)}}{12}\right) + \beta_{sk}\sin\left(\frac{2k\pi X_i^{(t)}}{12}\right)\right]. \quad (4.32)$$

Here β in the term $X_i^{(D)}\beta$ is used to capture the effect of the categorical demographic variables; $f_j(X_{ij}^{(C)})$ is estimated nonparametrically using cubic splines, capturing the effect of j^{th} continuous demographic variable. $f(X_i^{\text{loc}}) = f(X_{i,\text{lat}}^{\text{loc}}, f(X_{i,\text{lon}}^{\text{Loc}})$ is estimated by thin-plate splines, which is used to capture the geographic variations. The term β_t is used to capture time trend, and $\sum_{k=1}^{6}\left[\beta_{ck}\cos\left(\frac{2k\pi X_i^{(t)}}{12}\right) + \beta_{sk}\sin\left(\frac{2k\pi X_i^{(t)}}{12}\right)\right]$ is used to capture seasonal variation.

The `gam` function in `mgcv` package can be used to estimate the parameters and functions specified in the proposed GAM. In addition, model diagnoses and significance of each term can be obtained. A parsimonious model could be obtained for final interpretation.

4.3 Study of antibiotic overuse based on Kentucky Medicaid data

4.3.1 Background and data set

For the case study, we used the data from Kentucky Medicaid database from January 2014 to December 2016. Medicaid is a joint federal and state program that covers a wide range of health-related services and serves people who have limited financial resources and have diverse health care needs (Mainous III and Hueston, 1998). The Kentucky Medicaid data set contains information of beneficiaries who are insured by Kentucky Medicaid and have seen a doctor at least once in some

place in Kentucky. The data set contains three components: medical visit data file, pharmacy claim data file, and enrollment data file. Medical visit data file contains records of medical visits, diagnosis codes, information related to tests performed and medical providers. Pharmacy data file contains records of prescribed medication and pharmacy provider information, and date of services. Enrollment data file contains primarily demographic information of enrolled patients such as name, sex, age, race and foster care status. These data files can be linked together with patient identification number. In this study, we investigated the antibiotic overuse on children aged between 3 months to 17 years old and diagnosed with URI using Kentucky Medicaid data for years 2014-2016. The patients who met the above restrictions but were diagnosed with chronic and competing diseases are excluded. In order to appropriately identify the antibiotics overuse, we followed the restrictions provided by the healthcare effectiveness data and information set (HEDIS) (`http://www.ncqa.org/hedis-quality-measurement`). In addition, we include the health resource variable, the number of pediatricians per 10000 population in a county from the health resources and services administration data (`https://www.hrsa.gov/`). We also include the zip code level socio-economic covariates such as unemployment rate and percentage of poverty which are obtained from American FactFinder website `https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml`.

For the categorical covariates, we have demographic variables such as gender, race of the patient (white, black or other), medical region where the patient resides, and provider identification numbers. The continuous variables include age of the patient at visit, number of pediatricians per 10000 persons in a county where the patient resides, and percentage of poverty and unemployment rate in a zip code where the patient resides. The final data set includes 334254 visits of pediatric population of age between 3 months and 17 years, insured by Medicaid and diagnosed with URI but without other competing diseases.

We define an antibiotic prescription for URI as overuse if a patient is diagnosed with URI but without other concurrent chronic conditions and competing diseases. We hypothesize that the antibiotic overuse is associated with socioeconomic status and geographic areas. There are 8 different medicaid managed care organizations (MCO) regions in KY (Figure 4.1). We investigate whether there are variations among diffferent regions, even more granular, among different counties and zip code levels. As a secondary objective of this project, we also intend to identify the medical providers who contribute to the antibiotic overuse significantly.

4.3.2 Analysis and results

To understand the spatial pattern of zip code level antibiotics overuse in Kentucky, we plot the average antibiotic prescriptions per child for each zip code in Figure 4.2. Comparing Figure 4.2 to Figure 4.1, it is clear that medical regions 1 and 2 have higher prevalence of antibiotic overuse than other regions. Moreover, the variability of the overuse within each region across different zip codes supports our hypothesis that there are significant variation of antibiotic overuse across different zip code levels within the same region. We have also summarized the demographic and socio-economic variables in Table 4.1.

We have used `R` to perform the analyses. The popular function to fit GAM is `gam()` from the `mgcv` package. The function `bam()` from the same package is more suitable for handling large data sets. The estimated parameters obtained from the `bam()` output are presented in Table 4.2.

From Table 4.2, we can see that gender is not significant for antibiotic overuse. Race is significantly associated with antibiotic overuse and African American children are more prone to get antibiotic prescriptions. Patients from medical regions 1, 2, and 3 are more prone to get antibiotic overuse.

The complex relationships between continuous variables (e.g., age in months,

percentage of poverty level) and antibiotic overuse are shown in Figure 4.3. From Figure 4.3a, we can see that patients with age less than 50 months are more susceptible to get antibiotic overuse. Figure 4.3b indicates that patients from region with low percentage of poverty (i.e., rich areas) are more likely to receive antibiotic prescriptions for URI. On the other hand, the number of pediatricians per 10000 population at county level and the unemployment rate at zip code level are not associated with the antibiotic overuse.

Figure 4.4 demonstrates the temporal variation of antibiotic overuse. The steep downward slope in Figure 4.4a indicates that the antibiotic overuse has decreased significantly from 2014 through 2016. Figure 4.4b indicates that the antibiotic overuse takes a peak around June, which makes more sense since the season from April-June is the season of allergy, and children tend to be infected with diseases caused by air borne viruses. The overall trend in Figure 4.4c indicates that the antibiotic overuse has declined in Kentucky.

Table 4.3 provides the information of top 20 (ranked based on estimated odds ratio) health care providers and their observed percentages of antibiotic prescriptions to the children diagnosed with URI. From Table 4.3, it is clear that all of these providers prescribed antibiotics almost every time for the patients diagnosed with URI, indicating that more education and training may be needed for these providers.

4.4 Conclusion and discussion

In this project, we have proposed an approach using GAM to address the spatial and seasonal variability of the response. We used `bam()` function since we had a large number of observations.

There are some limitations related to this study. Since the population is only from the Kentucky Medicaid database, the findings obtained from the data analysis may not be generalized to the entire Kentucky population. We modeled the spatial

variation in the zip code level so that the variation of antibiotic overuse can be identified not only in medical regions but also within each zip code. However the data set had 31 zip codes which had small number (i.e., less than 10) of observations to depict the complete picture. Even the Medicaid data had its own flaws. The pharmacy claims data does not have a variable/indicator to associate with the medical claims. Hence it becomes challenging to ensure the linkage between the medication with the associated diagnosis. We have followed HEDIS guidelines to link medical claims with pharmacy claims assuming that the date of pharmacy visit should be within 3 days of the medical visit. This 3 day window is a subjective choice and might not ensure the perfect linking between medical claims and pharmacy claims data.

Despite the limitations of the model and the flaws of the data, the results obtained from the model portrays the reality. The descending overall trend of the antibiotic overuse implies that the awareness is increasing. On one hand, we see that the children are at high risk of getting unnecessary antibiotics prescription. On the other hand, we see that there are so many health care providers (Table 4.3) who prescribe unnecessary antibiotics prescriptions. This may be due to the fact that the antibiotic overuse prescription could come from the physicians who provide care for children but are not trained as pediatricians (Nyquist et al., 1998), or due to the unrealistic expectations and pressure from the patient family (Snow et al., 2001). In summary, the awareness is required to grow, education and training should target both physicians and the parents in order to improve the antibiotic prescribing practices (Nyquist et al., 1998).

4.5 Tables and Figures

Table 4.1: Summarized description of covariates across levels of antibiotic overuse

	No antibiotic overuse N=136938 (40.97%)	**Antibiotic overuse** N=197316 (59.03%)	**p value**
Continuous variables	**Mean (SD)**	**Mean (SD)**	
Age	78.48 (61.08)	77.96 (60.61)	0.016
Pediatrician count (per 10000)	1.33 (1.13)	1.01 (0.99)	<0.001
Unemployment rate	9.96 (5.25)	10.17 (5.83)	<0.001
Percentage of poverty	22.38 (9.69)	23.56 (9.84)	<0.001
Categorical variables	**N (%)**	**N (%)**	
Sex = Male	69190 (50.5%)	99329 (50.3%)	0.291
Race			<0.001
White	103053 (75.3%)	157988 (80.1%)	
Black	12591 (9.2%)	11707 (5.9%)	
Others	21294 (15.6%)	27621 (14.0%)	
Medical Regions			<0.001
Region 1	5702 (4.2%)	13436 (6.8%)	
Region 2	12164 (8.9%)	23147 (11.7%)	
Region 3	33093 (24.2%)	25572 (13.0%)	
Region 4	17793 (13.0%)	36562 (18.5%)	
Region 5	22986 (16.8%)	24589 (12.5%)	
Region 6	6565 (4.8%)	6960 (3.5%)	
Region 7	10225 (7.5%)	11641 (5.9%)	
Region 8	28410 (20.7%)	55409 (28.1%)	
Year			<0.001
2014	44753 (32.7%)	78264 (39.7%)	
2015	46560 (34.0%)	72868 (36.9%)	
2016	45625 (33.3%)	46184 (23.4%)	

Table 4.2: Estimated association between antibiotics overuse and sex, race, and medical regions

Variables	Odds Ratio	Lower CI	Upper CI	p-value
Sex				
Female	0.996	0.988	1.004	0.297
Male	1.004	0.996	1.012	0.297
Race				
Black	1.065	1.050	1.081	<0.001
Others	0.911	0.891	0.932	<0.001
White	1.030	1.012	1.048	0.001
Medical Regions				
Region 1	1.484	1.168	1.886	0.001
Region 2	1.259	1.100	1.441	0.001
Region 3	1.228	1.115	1.353	<0.001
Region 4	0.901	0.825	0.983	0.020
Region 5	0.786	0.721	0.857	<0.001
Region 6	0.778	0.659	0.919	0.003
Region 7	1.011	0.901	1.135	0.853
Region 8	0.782	0.707	0.866	<0.001

Table 4.3: The top 20 health care providers who prescribed antibiotics for patients diagnosed with URI

ID	OR	Observed %	Frequency
*******818	404.90	0.995	193
*******925	234.73	0.991	107
*******445	146.22	0.980	50
*******488	137.44	0.975	81
*******093	135.76	0.982	283
*******865	131.79	0.981	53
*******301	117.15	0.976	85
*******154	104.26	0.990	99
*******272	99.01	0.967	30
*******538	93.62	0.969	32
*******786	90.56	0.986	73
*******159	89.36	0.989	92
*******303	83.08	0.972	692
*******791	73.65	0.982	54
*******367	71.49	0.974	39
*******306	70.04	0.973	73
*******707	68.46	0.959	221
*******115	67.85	0.973	74
*******630	66.37	0.962	26
*******700	66.16	0.966	59

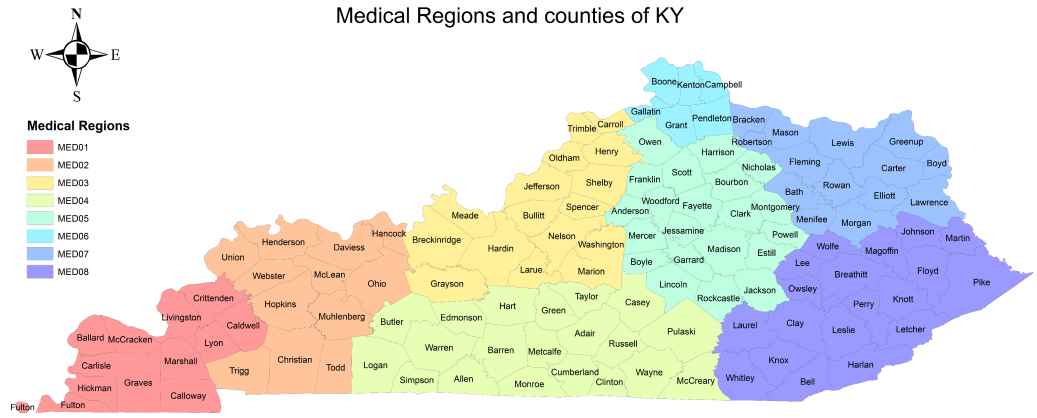


Figure 4.1: KY Medicaid MCO regions (Image taken from Marton et al. (2016))

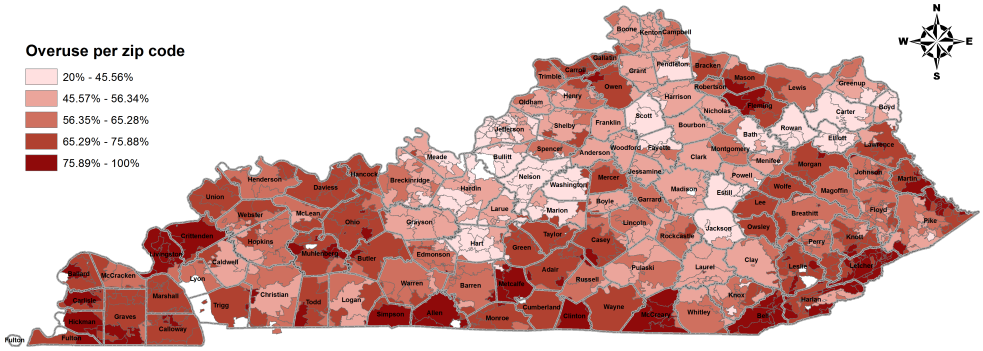
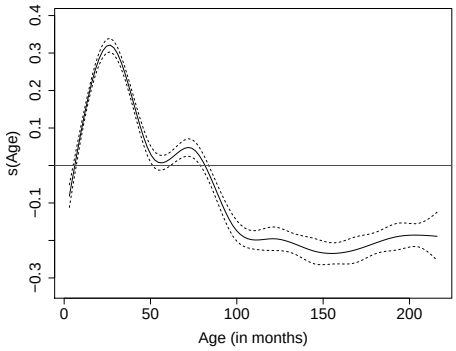
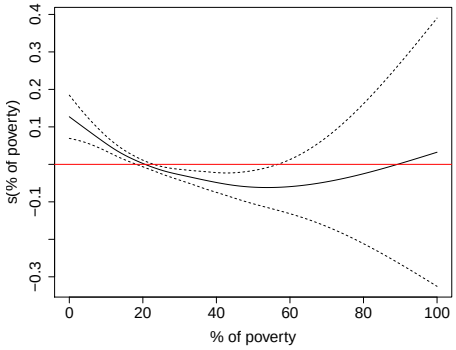


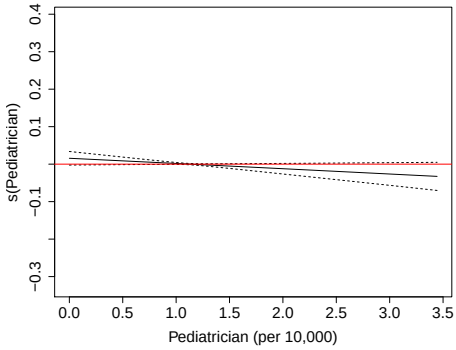
Figure 4.2: Fraction of antibiotic overuse across different zip codes for children diagnosed with URI based on 2014-2016 Kentucky Medicaid data



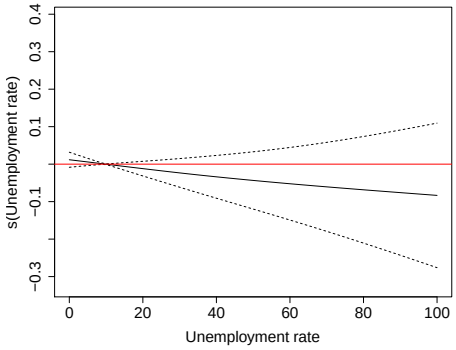
(a) Antibiotic overuse vs age (p <0.001)



(b) Antibiotic overuse vs poverty level (p <0.001)

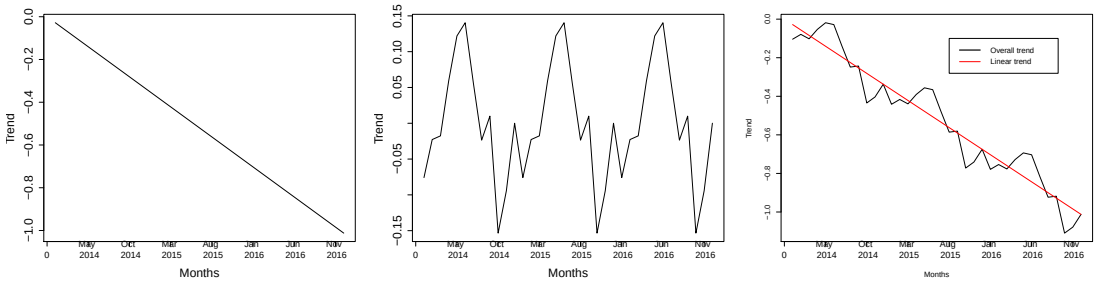


(c) Antibiotic overuse vs pediatricians counts (p=0.085)



(d) Antibiotic overuse vs unemployment rate (p=0.279)

Figure 4.3: The complex association between antibiotic overuse and (a) age in months; (b) percentage of poverty level; (c) number of pediatricians; and (d) unemployment rate



(a) Time trend (b) Seasonal variation (c) Overall time effect

Figure 4.4: Illustration of time trend and seasonal variation

REFERENCES

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., and Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59(5):967–985.

Adams, R., Harrison, J., Scott, P., et al. (1969). The development of cadmium-induced proteinuria, impaired renal function, and osteomalacia in alkaline battery workers. *Quarterly Journal of Medicine*, 38(152):425–43.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley-Interscience, Hoboken, New Jersey.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley & Sons, Hoboken, New Jersey.

Agresti, A. and Kateri, M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics*, 73(1):214–219.

Aktekin, T. and Musal, M. (2015). Analysis of income inequality measures on Human Immunodeficiency Virus mortality: A spatiotemporal Bayesian perspective. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(2):383–403.

Alumran, A., Hurst, C., and Hou, X.-Y. (2012). Antibiotics overuse in children with upper respiratory tract infections in saudi arabia: Risk factors and potential interventions. *Clinical Medicine and Diagnostics*, 1(1):8–16.

Andersen, R. M. (1995). Revisiting the behavioral model and access to medical care: Does it matter? *Journal of Health and Social Behavior*, 36(1):1–10.

Angers, J.-F. and Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics & Data Analysis*, 42(1):37 – 46.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679.

Braun, W. J. and Huang, L.-S. (2005). Kernel spline regression. *Canadian Journal of Statistics*, 33(2):259–278.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37.

Choo-Wosoba, H., Gaskins, J., Levy, S., and Datta, S. (2018). A Bayesian approach for analyzing zero-inflated clustered count data with dispersion. *Statistics in Medicine*, 37(5):801–812. sim.7541.

Dobbie, M. J. and Welsh, A. (2001). Theory & methods: Modelling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics*, 43(4):431–444.

Fang, X. and Chan, K.-S. (2014). Additive models with spatio-temporal data. *Environmental and Ecological Statistics*, 22(1):61–86.

Garçon, G., Leleu, B., Marez, T., Zerimech, F., Haguenoer, J.-M., Furon, D., and Shirali, P. (2007). Biomonitoring of the adverse effects induced by the chronic exposure to lead and cadmium on kidney function: usefulness of alpha-glutathione s-transferase. *Science of the Total Environment*, 377(2-3):165–172.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1(3):515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Giberson, A., Vaziri, N. D., Mirahamadi, K., and Rosen, S. (1976). Hemodialysis of acute arsenic intoxication with transient renal failure. *Archives of internal medicine*, 136(11):1303–1304.

Gonzales, R., Steiner, J. F., and Sande, M. A. (1997). Antibiotic prescribing for adults with colds, upper respiratory tract infections, and bronchitis by ambulatory care physicians. *JAMA*, 278(11):901–904.

Guerrero, V. M. and Johnson, R. A. (1982). Use of the box-cox transformation with binary response models. *Biometrika*, 69(2):309–314.

Guo, S. (2010). *Propensity score analysis : Statistical Methods and Applications*. Sage Publications, Thousand Oaks, California.

Hall, D. B. (2000). Zero-inflated Poisson and Binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039.

Hall, D. B. and Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling*, 4(3):161–180.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294.

Hicks, L. A., Bartoces, M. G., Roberts, R. M., Suda, K. J., Hunkler, R. J., Taylor, Jr, T. H., and Schrag, S. J. (2015). Us outpatient antibiotic prescribing variation according to geography, patient population, and provider specialty in 2011. *Clinical Infectious Diseases*, 60(9):1308–1316.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hur, K., Hedeker, D., Henderson, W., Khuri, S., and Daley, J. (2002). Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology*, 3(1):5–20.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Klotz, J. H. (1966). The wilcoxon, ties, and the computer. *Journal of the American Statistical Association*, 61(315):772–787.

Kong, M., Cambon, A., and Smith, M. J. (2012). Extended logistic regression model for studies with interrupted events, seasonal trend, and serial correlation. *Communications in Statistics-Theory and Methods*, 41(19):3528–3543.

Kong, M., Xu, S., Levy, S. M., and Datta, S. (2015). GEE type inference for clustered zero-inflated Negative Binomial regression with application to dental caries. *Computational Statistics & Data Analysis*, 85:54 – 66.

Kruskal, W. H. (1957). Historical notes on the wilcoxon unpaired two-sample test. *Journal of the American Statistical Association*, 52(279):356–360.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K., and McLachlan, G. J. (2006). Multilevel zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, 15(1):47–61. PMID: 16477948.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Levey, A. S. and Coresh, J. (2012). Chronic kidney disease. *The Lancet*, 379(9811):165–180.

Levey, A. S., Coresh, J., Bolton, K., Culleton, B., Harvey, K. S., Ikizler, T. A., Johnson, C. A., Kausz, A., Kimmel, P. L., Kusek, J., et al. (2002). K/doqi clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases*, 39(2 SUPPL. 1).

Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T., et al. (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, 150(9):604–612.

Lim, H. K., Song, J., and Jung, B. C. (2013). Score tests for zero-inflation and overdispersion in two-level count data. *Computational Statistics & Data Analysis*, 61:67 – 82.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1):35 – 46.

Mainous III, A. G. and Hueston, W. J. (1998). The cost of antibiotics in treating upper respiratory tract infections in a medicaid population. *Archives of family medicine*, 7(1):45.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Marton, J., Talbert, J., Palmer, A., Howell, E., Costich, J., Wissoker, D., and Kenney, G. M. (2016). Medicaid managed care in Kentucky.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.

McCullagh, P. (1989). *Generalized linear models*. Chapman and Hall, London New York.

Meliker, J. R., Slotnick, M. J., AvRuskin, G. A., Kaufmann, A., Jacquez, G. M., and Nriagu, J. O. (2005). Improving exposure assessment in environmental epidemiology: Application of spatio-temporal visualization tools. *Journal of Geographical Systems*, 7(1):49–66.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341 – 365.

Musal, M. and Aktekin, T. (2013). Bayesian spatial modeling of HIV mortality via zero-inflated Poisson models. *Statistics in Medicine*, 32(2):267–281.

Nelder, J. A. and Baker, R. J. (1972). *Generalized linear models*. Wiley Online Library.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(1):3–48.

Nyquist, A. C., Gonzales, R., Steiner, J. F., and Sande, M. A. (1998). Antibiotic prescribing for children with colds, upper respiratory tract infections, and bronchitis. *JAMA*, 279(11):875–877.

Peuquet, D. J. and Duan, N. (1995). An event-based spatiotemporal data model

(estdm) for temporal analysis of geographical data. *International journal of geographical information systems*, 9(1):7–24.

Ratnaike, R. N. (2003). Acute and chronic arsenic toxicity. *Postgraduate Medical Journal*, 79(933):391–396.

Ridout, M., Hinde, J., and DeméAtrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated Negative Binomial alternatives. *Biometrics*, 57(1):219–223.

Robins, J. (2000). Marginal structual models and causal inference in epidemiology. *Epidemiology*, 11:95–134.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Ryu, E. and Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, 27(10):1703–1717.

Satten, G. A., Kong, M., and Datta, S. (2018). Multi-sample adjusted u-statistics that account for confounding covariates. *Statistics in Medicine (In Press)*.

Shankar, V., Milton, J., and Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention*, 29(6):829 – 837.

Snow, V., Mottur-Pilson, C., Gonzales, R., and for the American College of Physicians–American Society of Internal Medicine* (2001). Principles of appropriate antibiotic use for treatment of nonspecific upper respiratory tract infections in adults. *Annals of Internal Medicine*, 134(6):487–489.

Soyka, L. F., Robinson, D. S., Lachant, N., and Monaco, J. (1975). The misuse of antibiotics for treatment of upper respiratory tract infections in children. *Pediatrics*, 55(4):552–556.

Vargha, A. and Delaney, H. D. (1998). The kruskal-wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, 23(2):170–192.

Vaziri, N. D., Upham, T., and Barton, C. H. (1980). Hemodialysis clearance of arsenic. *Clinical Toxicology*, 17(3):451–456. PMID: 7449358.

Wang, K., Yau, K. K., and Lee, A. H. (2002). A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. *Computer Methods and Programs in Biomedicine*, 68(3):195 – 203.

Watson, R. L., Dowell, S. F., Jayaraman, M., Keyserling, H., Kolczak, M., and Schwartz, B. (1999). Antimicrobial use for pediatric upper respiratory infections: reported practice, actual practice, and parent beliefs. *Pediatrics*, 104(6):1251–1257.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Xie, F.-C., Lin, J.-G., and Wei, B.-C. (2014). Bayesian zero-inflated generalized Poisson regression model: Estimation and case influence diagnostics. *Journal of Applied Statistics*, 41(6):1383–1392.

Xu, J., Kockelman, K. M., and Wang, Y. (2014). Modeling crash and fatality counts along mainlines and frontage roads across Texas: The roles of design, the built environment, and weather. In *Transportation Research Board, 93rd Annual Meeting*, volume 22, page 24.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065.

Yau, K. K. W., Wang, K., and Lee, A. H. (2003). Zero-inflated Negative Binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452.

Zhang, C., Chen, N., and Zhang, L. (2016). Time series of multivariate zero-inflated Poisson counts. In *Industrial Engineering and Engineering Management (IEEM), 2016 IEEE International Conference on*, pages 1365–1369. IEEE.

Zhang, Y., Steinman, M. A., and Kaplan, C. M. (2012). Geographic variation in outpatient antibiotic prescribing among older adults. *Archives of internal medicine*, 172(19):1465–1471.

APPENDIX

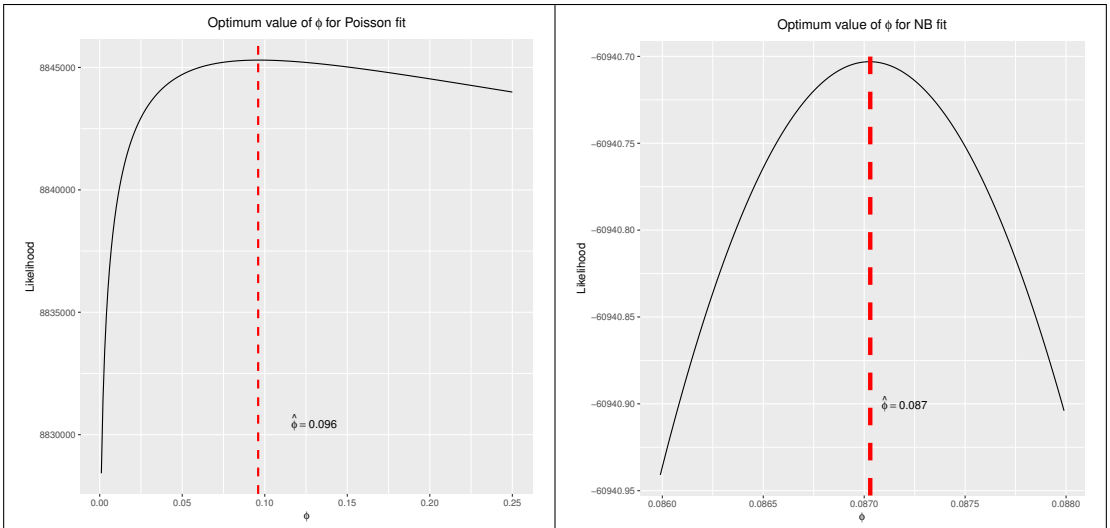
Appendix A

This section includes the additional figures and tables in Chapter 2.

A.1 Estimate and variance of ϕ

Estimate of ϕ

As discussed in Section 2.2.2, we have used multistage approach to estimate the dispersion parameter ϕ. Figure A1.1a illustrates the approximated marginal posterior function (i.e., $\hat{l}(\phi|\mathbf{Y})$) of ϕ based on $\boldsymbol{\theta}_-$ obtained from Poisson model and observed data, where the maximum of ϕ is 0.096, represented by the dotted line. Figure A1.1b depicts the marginal posterior function of ϕ based $\boldsymbol{\theta}_-$ obtained from the NB model at the final estimate of ϕ.



(a) Initial estimate of $\hat{\phi}$ based on the approximated marginal posterior function of ϕ

(b) Final estimate of $\hat{\phi}$ based on the marginal posterior function of ϕ

Figure A1.1: Estimation of ϕ using two-stage approach

Variance of ϕ

To obtain the Wald-style credible interval, we need the observed information I_{obs} for ϕ from the posterior $\pi(\phi|\mathbf{Y})$. This is obtained by evaluating $-\frac{d^2}{d\phi^2}\log\pi(\phi|\mathbf{Y})$ at the final estimate $\hat{\phi}$ and approximating the form of $\pi(\phi|\mathbf{Y})$ using equations (2.9) and (2.10). We have

$$\begin{aligned}\frac{d}{d\phi}\log\hat{\pi}(\phi|\mathbf{Y}) &\approx \frac{d}{d\phi}\left\{\log\pi(\phi) - \log m(\mathbf{Y}) - \log\left[\frac{1}{G}\sum_g e^{-\sum_{i,j,k} l(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})}\right]\right\} \\ &= \frac{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})} l'(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}},\end{aligned}$$

where $l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y}) = \sum_{i,j,k} l(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})$ and $l'(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y}) = \frac{d}{d\phi}\sum_{i,j,k} l(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})$.

Then the observed information (I_{obs}) for ϕ can be obtained as

$$
\begin{aligned}
-I_{obs} &= \frac{d^2}{d\phi^2}\log\hat{\pi}(\phi|\mathbf{Y})\Big|_{\phi=\hat{\phi}} = \frac{d}{d\phi}\left(\frac{d\log\hat{\pi}(\phi|\mathbf{Y})}{d\phi}\right)\Big|_{\phi=\hat{\phi}} \\
&\approx \frac{d}{d\phi}\left(\frac{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})} l'(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}}\right)\Big|_{\phi=\hat{\phi}} \\
&= \frac{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}\left[l''(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y}) - \{l'(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})\}^2\right]}{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}}\Bigg|_{\phi=\hat{\phi}} \\
&+ \left\{\frac{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})} l'(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}{\sum_g e^{-l(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y})}}\right\}^2\Bigg|_{\phi=\hat{\phi}},
\end{aligned}
$$

where $l''(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y}) = \frac{d^2}{d\phi^2}\sum_{i,j,k} l(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})$.

The variance of the overdispersion parameter ϕ is estimated by $\hat{s}_\phi^2 = 1/I_{obs}$, and the 95% credible interval is constructed by $\hat{\phi} \pm 1.96\sqrt{\hat{s}_\phi^2}$.

The mathematical forms of $l'(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})$ and $l''(\boldsymbol{\theta}_-^{(g)},\phi|y_{ijk})$ are given as

$$
\begin{aligned}
l'(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y}) &= \sum_i\sum_j\sum_k \{1 - I(y_{ijk}=0)\}\Bigg[-\Gamma_2\left(y_{ijk}+\frac{1}{\phi}\right)\frac{1}{\phi^2} - \Gamma_2\left(\frac{1}{\phi}\right)\frac{1}{\phi^2} - \\
&\qquad \frac{u_1u_2}{f_0} + \frac{y_{ijk}}{\phi} - \frac{y_{ijk}\mu_{ijk}}{1+\phi\mu_{ijk}} - u_1\Bigg] \\
l''(\boldsymbol{\theta}_-^{(g)},\phi|\mathbf{Y}) &= \sum_i\sum_j\sum_k \{1 - I(y_{ijk}=0)\}[f_1 - f_2 - f_3 + f_4],
\end{aligned}
$$

Here,

$$
\begin{aligned}
u_1 &= \frac{\mu_{ijk}}{\phi\left(1+\phi\mu_{ijk}\right)} - \frac{1}{\phi^2}\log\left(1+\phi\mu_{ijk}\right),\\
u_2 &= \frac{1}{\left(1+\phi\mu_{ijk}\right)^{\frac{1}{\phi}}},\\
f_1 &= \Gamma_3\left(y_{ijk}+\frac{1}{\phi}\right)\frac{1}{\phi^4} + \Gamma_2\left(y_{ijk}+\frac{1}{\phi}\right)\frac{2}{\phi^3},\\
f_2 &= \Gamma_3\left(\frac{1}{\phi}\right)\frac{1}{\phi^4} + \Gamma_2\left(\frac{1}{\phi}\right)\frac{2}{\phi^3},\\
f_3 &= \nu_1+\nu_2,\\
f_4 &= y_{ijk}\log(\phi y_{ijk}) - y_{ijk}\log(1+\phi y_{ijk}) - \frac{1}{\phi}\log(1+\phi y_{ijk}),\\
\nu_1 &= \frac{1}{f_0}\left[\frac{1}{\left(1+\phi\mu_{ijk}\right)^{\frac{1}{\phi}}}\left\{-\frac{\mu_{ijk}(1+2\phi\mu_{ijk})}{\{\phi(1+\phi\mu_{ijk})\}^2} - \frac{\frac{\phi^2\mu_{ijk}}{1+\phi\mu_{ijk}} - 2\phi\log(1+\phi\mu_{ijk})}{\phi^4}\right\}+\right.\\
&\left.\frac{\left\{\frac{\mu_{ijk}}{\phi(1+\phi\mu_{ijk})} - \frac{\log(1+\phi\mu_{ijk})}{\phi^2}\right\}^2}{\left(1+\phi\mu_{ijk}\right)^{\frac{1}{\phi}}}\right],\\
\nu_2 &= \left\{\frac{\frac{\mu_{ijk}}{\phi(1+\phi\mu_{ijk})} - \frac{\log(1+\phi-\mu_{ijk})}{\phi^2}}{f_0\left(1+\phi\mu_{ijk}\right)^{\frac{1}{\phi}}}\right\}^2,\\
f_0 &= \left(1-\frac{1}{\left(1+\phi\mu_{ijk}\right)^{\frac{1}{\phi}}}\right).
\end{aligned}
$$

With the usual notation, $\Gamma(x)$ is the Gamma function, $\Gamma_2(x) = \frac{d}{dx}\log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the double Gamma function, and $\Gamma_3(x) = \frac{d^2}{dx^2}\log(\Gamma(x))$ is the triple Gamma function.

A.2 Model checking and validation

A common method for Bayesian model validation is to compare replicated data sets drawn from the posterior predictive distribution to the real data (Gelman et al., 2013). Let $\boldsymbol{y}^{(r)}$ be the r^{th} replicated data set, $r = 1, \ldots, R$, which comes from $p(\boldsymbol{y}^{rep} \mid \boldsymbol{y}) = \int p(\boldsymbol{y}^{rep} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}$. If the chosen model $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ appropriately describes the data, then the replicated data set $\boldsymbol{y}^{(r)}$ will be similar to the real data.

To investigate this, the analyst will select a typically univariate summary statistic $t(\boldsymbol{y})$ that captures an important feature of the model and compare the distribution of $t(\boldsymbol{Y}^{rep})$ to the observed value $t(\boldsymbol{y})$. If $t(\boldsymbol{y})$ is near the mode (or within the 95% interval) of $t(\boldsymbol{Y}^{rep})$, then we conclude the data model provides a good fit to the real data (at least in terms of the features described by the choice of $t(\cdot)$).

To that end, we select 6 summary statistics $t_k(\boldsymbol{y})$, $k = 1, \ldots, 6$, to represent key features of the data and our model structure. As shown in Section 2.3.2, population size plays a dominant role in the number of primary care physicians. Therefore, we base some of our summary statistics on a population standardized version of the data $\tilde{\boldsymbol{y}}$ by letting $\tilde{y}_{ijk}$ be y_{ijk} divided by the year k population of county j of state i. This leads to an overall reduction in the noise of the data sets and better allows us to inspect the features of the model. Of the six summary statistics, the first three are chosen to assess the goodness of the overall fit of the model to the data and the final three are chosen to validate our spatial/temporal dependence model. The chosen summary statistics are as follows:

1. $t_1(\boldsymbol{y}) = t_1(\tilde{\boldsymbol{y}}) = \overline{\tilde{y}_{\cdots}}$ is the mean of the population-standardized primary care physicians counts. That is the average per capita number of primary care physicians across the nation.

2. $t_2(\boldsymbol{y})$ consider the median of the (unstandardized) y_{ijk}.

3. $t_3(\boldsymbol{y})$ is the proportion of responses y_{ijk} that are zero.

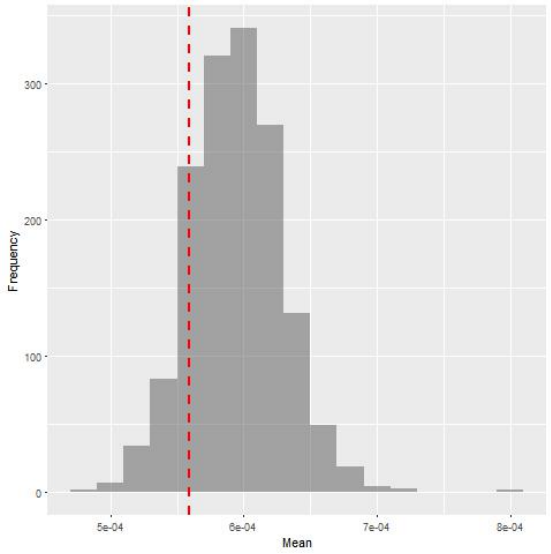
4. $t_4(\boldsymbol{y}) = t_4(\tilde{\boldsymbol{y}})$ compares the similarity of the per-capita counts within states. $t_4(\boldsymbol{y})$ is the ratio of the variance of $\overline{\tilde{y}_{i\cdot\cdot}}$ (average per capita count for state i across all counties and all years) to the overall variance of the $\tilde{y}_{ijk}$.

5. $t_5(\boldsymbol{y}) = t_5(\tilde{\boldsymbol{y}})$ considers the portion of the variability explained by the spatial component. For each county j we determine the knot $l = l(i, j)$ that is closest

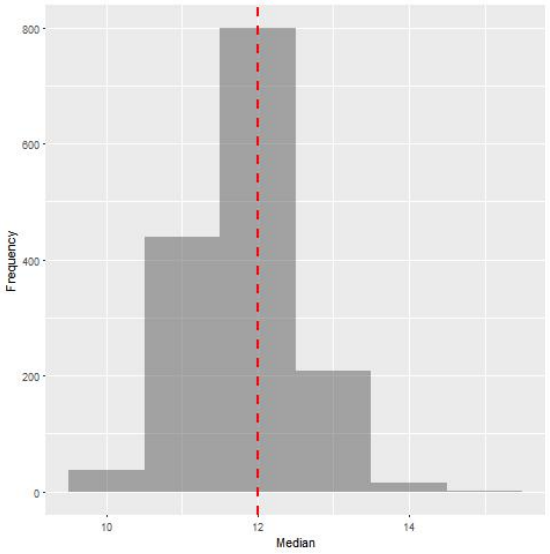
to it, and let z_l be the average of $\tilde{y}_{ijk}$ of all counties with $l = l(i, j)$. The statistic $t_5(\boldsymbol{y})$ is the ratio of the variance of z_l to the variance of $\tilde{y}_{ijk}$.

6. $t_6(\boldsymbol{y}) = t_6(\tilde{\boldsymbol{y}})$ is the same as $t_4(\boldsymbol{y})$ except that we consider $\overline{\tilde{y}_{i\cdot k}}$, the average per capita count for state i in year k averaged across all counties.

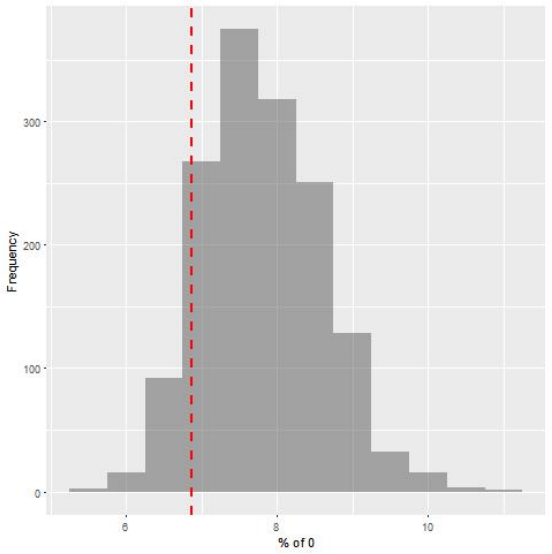
We generate $R = 1500$ replicated data sets and plot the histograms of $t_k(\boldsymbol{Y}^{rep})$ in Figure A1.2. The corresponding value $t_k(\boldsymbol{y})$ for the true data is shown in the dashed line. All six values of $t_k(\boldsymbol{y})$ are contained in the 95% intervals of $t_k(\boldsymbol{Y}^{rep})$, and with the possible exception of $t_6(\boldsymbol{y})$, they are close to the mode. Therefore, we conclude that our model framework is consistent with the HPSA data on which it is applied.



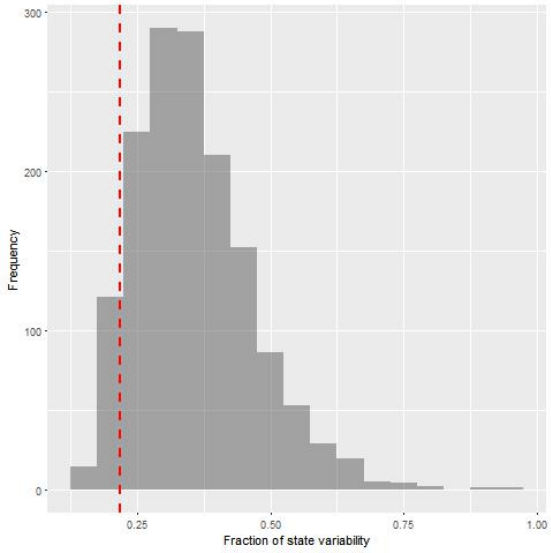
(a) Posterior predictive distribution for $t_1(\boldsymbol{y})$ (standardized)



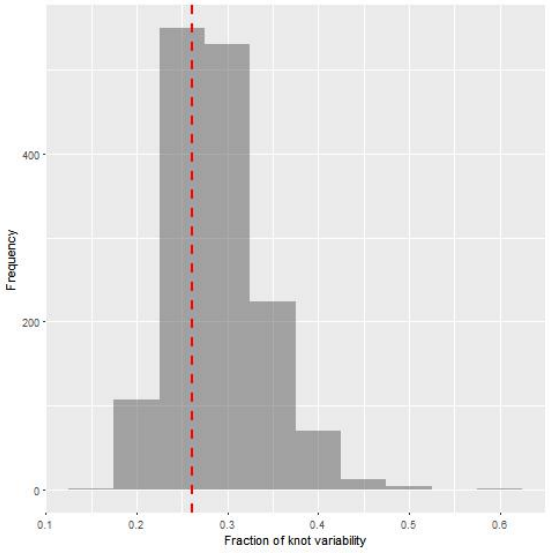
(b) Posterior predictive distribution for $t_2(\boldsymbol{y})$



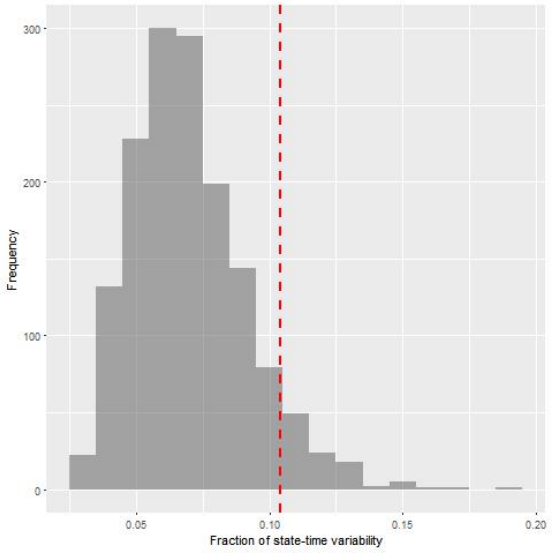
(c) Posterior predictive distribution for $t_3(\boldsymbol{y})$



(d) Posterior predictive distribution for $t_4(\boldsymbol{y})$ (standardized)



(e) Posterior predictive distribution for $t_5(\boldsymbol{y})$ (standardized)



(f) Posterior predictive distribution for $t_6(\boldsymbol{y})$ (standardized)

Figure A1.2: Illustration of different kinds of Posterior predictive plots

A.3 Additional simulation results

Recall from Section 2.4, we explored two additional simulations studies. In Table A1.1, the true data generating model is NB-ST, and in Table A1.2, the true model is POI-SP-ST. As shown in Table A1.1, the performance of our proposed NB-SP-ST model and the true NB-ST model is mostly indistinguishable. There is some evidence that in the count model NB-SP-ST may be better than NB-ST. In Table A1.2 where the true model is Poisson, our NB-SP-ST model effectively estimates ϕ to be the limiting value near 0 and achieves estimation loss equivalently to, or slightly better than, the true POI-SP-ST model.

Table A1.1: Simulation results when the underlying model was NB-ST, but the data was fitted with different models

		True	**NB-SP-ST**			**NB-ST**			**POI-SP-ST**			**POI-ST**		
		Value	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$
Binary Model	β_0	2.2	2.215	0.015	3.211	2.222	0.022	3.123	2.210	0.010	3.261	2.210	0.010	3.087
	β_1	1.3	1.303	0.003	0.496	1.304	0.004	0.486	1.300	0.000	0.545	1.300	0.000	0.512
	β_2	0	0.011	0.011	0.260	0.011	0.011	0.264	0.012	0.012	0.261	0.011	0.011	0.261
	β_3	-0.1	-0.076	0.024	0.732	-0.077	0.023	0.728	-0.076	0.024	0.769	-0.076	0.024	0.745
	β_4	0	-0.002	-0.002	0.121	-0.002	-0.002	0.123	-0.002	-0.002	0.132	-0.003	-0.003	0.133
	β_5	0.1	0.100	0.000	0.256	0.100	0.000	0.256	0.101	0.001	0.265	0.100	0.000	0.255
	β_6	0	-0.012	-0.012	0.249	-0.012	-0.012	0.249	-0.014	-0.014	0.255	-0.013	-0.013	0.253
	β_7	0	-0.005	-0.005	0.190	-0.005	-0.005	0.191	-0.005	-0.005	0.201	-0.006	-0.006	0.202
	β_8	0	-0.005	-0.005	0.355	-0.005	-0.005	0.355	-0.005	-0.005	0.360	-0.005	-0.005	0.344
	β_9	0.1	0.109	0.009	0.236	0.109	0.009	0.232	0.108	0.008	0.252	0.109	0.009	0.238
	β_{10}	0.1	0.097	-0.003	0.136	0.097	-0.003	0.136	0.096	-0.004	0.138	0.097	-0.003	0.142
	$t1_1$	0.05	0.032	-0.018	2.197	0.030	-0.020	2.164	0.036	-0.014	2.167	0.033	-0.017	2.215
	$t1_2$	0.1	0.110	0.010	1.427	0.107	0.007	1.433	0.110	0.010	1.473	0.108	0.008	1.449
	$t1_3$	0.15	0.136	-0.014	1.625	0.133	-0.017	1.639	0.138	-0.012	1.647	0.133	-0.017	1.684
	$t1_4$	0.2	0.208	0.008	1.659	0.203	0.003	1.731	0.206	0.006	1.682	0.204	0.004	1.675
	$t1_5$	0.25	0.244	-0.006	1.504	0.243	-0.007	1.511	0.242	-0.008	1.430	0.239	-0.011	1.444
	Loss*				14.656			**14.622**			14.839			14.640
Count Model	γ_0	2.2	2.198	-0.002	0.199	2.200	0.000	0.202	2.204	0.004	0.239	2.203	0.003	0.216
	γ_1	1.5	1.502	0.002	0.007	1.502	0.002	0.008	1.497	-0.003	0.014	1.498	-0.002	0.014
	γ_2	0	0.000	0.000	0.008	0.000	0.000	0.008	-0.001	-0.001	0.021	-0.001	-0.001	0.021
	γ_3	-0.9	-0.903	-0.003	0.018	-0.903	-0.003	0.018	-0.900	0.000	0.028	-0.901	-0.001	0.030
	γ_4	0	0.000	0.000	0.009	0.000	0.000	0.009	0.001	0.001	0.010	0.001	0.001	0.011
	γ_5	0.2	0.200	0.000	0.005	0.200	0.000	0.005	0.199	-0.001	0.008	0.199	-0.001	0.007
	γ_6	0	0.000	0.000	0.012	0.000	0.000	0.011	-0.001	-0.001	0.018	-0.001	-0.001	0.018
	γ_7	0.1	0.100	0.000	0.009	0.100	0.000	0.009	0.100	0.000	0.021	0.100	0.000	0.020
	γ_8	-0.3	-0.300	0.000	0.012	-0.300	0.000	0.012	-0.300	0.000	0.022	-0.300	0.000	0.025
	γ_9	0.2	0.200	0.000	0.009	0.200	0.000	0.009	0.198	-0.002	0.013	0.198	-0.002	0.015
	γ_{10}	0.2	0.200	0.000	0.011	0.200	0.000	0.011	0.200	0.000	0.018	0.201	0.001	0.019
	$t2_1$	0.025	0.025	0.000	0.067	0.024	-0.001	0.072	0.024	-0.001	0.119	0.023	-0.002	0.115
	$t2_2$	0.05	0.052	0.002	0.066	0.052	0.002	0.067	0.053	0.003	0.100	0.052	0.002	0.098
	$t2_3$	0.075	0.073	-0.002	0.082	0.073	-0.002	0.082	0.072	-0.003	0.106	0.070	-0.005	0.106
	$t2_4$	0.1	0.098	-0.002	0.058	0.097	-0.003	0.062	0.098	-0.002	0.083	0.098	-0.002	0.082
	$t2_5$	0.125	0.124	-0.001	0.056	0.124	-0.001	0.061	0.122	-0.003	0.079	0.121	-0.004	0.075
	Loss*				**0.627**			0.645			0.897			0.872
	ϕ	0.096	0.097	0.001	**0.003**	0.097	0.001	**0.003**	-	-	-	-	-	-
	a	3.5	3.377	-0.123	10.682	3.378	-0.122	**10.675**	3.059	-0.441	31.483	3.155	-0.345	23.917

Loss*: The sum of squared errors for the fixed effect coefficients

Table A1.2: Simulation results when the underlying model was POI-SP-ST, but the data was fitted with different models

		True	**NB-SP-ST**			**NB-ST**			**POI-SP-ST**			**POI-ST**		
		Value	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$	Est	Bias	$MSE \times 100$
Binary Model	β_0	2.2	2.236	0.036	3.512	2.138	-0.062	5.978	2.236	0.036	3.576	2.143	-0.057	5.389
	β_1	1.3	1.311	0.011	0.460	1.247	-0.053	1.150	1.310	0.010	0.454	1.245	-0.055	1.214
	β_2	0	0.009	0.009	0.333	0.008	0.008	0.594	0.009	0.009	0.331	0.008	0.008	0.613
	β_3	-0.1	-0.077	0.023	0.806	-0.055	0.045	1.084	-0.076	0.024	0.807	-0.053	0.047	1.110
	β_4	0	-0.006	-0.006	0.110	-0.011	-0.011	0.306	-0.005	-0.005	0.109	-0.011	-0.011	0.328
	β_5	0.1	0.103	0.003	0.245	0.088	-0.012	0.647	0.103	0.003	0.248	0.089	-0.011	0.656
	β_6	0	-0.009	-0.009	0.246	-0.014	-0.014	0.451	-0.009	-0.009	0.244	-0.014	-0.014	0.449
	β_7	0	-0.006	-0.006	0.226	-0.014	-0.014	0.464	-0.006	-0.006	0.229	-0.015	-0.015	0.490
	β_8	0	0.000	0.000	0.370	-0.013	-0.013	1.179	0.000	0.000	0.376	-0.012	-0.012	1.276
	β_9	0.1	0.112	0.012	0.282	0.112	0.012	0.430	0.112	0.012	0.280	0.113	0.013	0.454
	β_{10}	0.1	0.097	-0.003	0.167	0.088	-0.012	0.416	0.097	-0.003	0.170	0.088	-0.012	0.426
	$t1_1$	0.05	0.025	-0.025	2.057	0.020	-0.030	2.645	0.027	-0.023	2.093	0.019	-0.031	2.592
	$t1_2$	0.1	0.097	-0.003	1.835	0.082	-0.018	2.025	0.099	-0.001	1.902	0.084	-0.016	1.878
	$t1_3$	0.15	0.138	-0.012	1.627	0.122	-0.028	2.079	0.140	-0.010	1.591	0.124	-0.026	2.028
	$t1_4$	0.2	0.195	-0.005	1.764	0.183	-0.017	1.797	0.197	-0.003	1.770	0.184	-0.016	1.813
	$t1_5$	0.25	0.246	-0.004	2.019	0.233	-0.017	1.761	0.250	0.000	1.989	0.234	-0.016	1.740
	Loss*				**16.059**			23.006			16.170			22.455
Count Model	γ_0	2.2	2.203	0.003	0.202	2.226	0.026	0.482	2.203	0.003	0.205	2.229	0.029	0.475
	γ_1	1.5	1.500	0.000	0.004	1.499	-0.001	0.036	1.500	0.000	0.004	1.498	-0.002	0.047
	γ_2	0	0.001	0.001	0.005	0.007	0.007	0.041	0.001	0.001	0.005	0.006	0.006	0.049
	γ_3	-0.9	-0.900	0.000	0.006	-0.901	-0.001	0.059	-0.900	0.000	0.006	-0.901	-0.001	0.079
	γ_4	0	-0.001	-0.001	0.004	-0.001	-0.001	0.031	-0.001	-0.001	0.004	-0.002	-0.002	0.040
	γ_5	0.2	0.200	0.000	0.003	0.197	-0.003	0.026	0.200	0.000	0.003	0.197	-0.003	0.038
	γ_6	0	-0.001	-0.001	0.004	0.002	0.002	0.031	-0.001	-0.001	0.004	0.003	0.003	0.036
	γ_7	0.1	0.102	0.002	0.003	0.099	-0.001	0.023	0.101	0.001	0.003	0.099	-0.001	0.031
	γ_8	-0.3	-0.298	0.002	0.005	-0.299	0.001	0.036	-0.298	0.002	0.005	-0.296	0.004	0.047
	γ_9	0.2	0.200	0.000	0.003	0.202	0.002	0.028	0.200	0.000	0.003	0.202	0.002	0.048
	γ_{10}	0.2	0.200	0.000	0.003	0.197	-0.003	0.025	0.200	0.000	0.003	0.196	-0.004	0.034
	$t2_1$	0.025	0.023	-0.002	0.045	0.024	-0.001	0.069	0.024	-0.001	0.047	0.024	-0.001	0.078
	$t2_2$	0.05	0.050	0.000	0.043	0.050	0.000	0.058	0.051	0.001	0.046	0.051	0.001	0.064
	$t2_3$	0.075	0.076	0.001	0.039	0.074	-0.001	0.053	0.076	0.001	0.040	0.075	0.000	0.061
	$t2_4$	0.1	0.097	-0.003	0.042	0.097	-0.003	0.051	0.098	-0.002	0.045	0.097	-0.003	0.049
	$t2_5$	0.125	0.123	-0.002	0.036	0.121	-0.004	0.042	0.124	-0.001	0.037	0.122	-0.003	0.041
	Loss*				**0.449**			1.092			0.461			1.214
	ϕ	0	0.001	0.001	$<$ **0.001**	0.024	0.024	0.065	-	-	-	-	-	-
	a	3.5	3.488	-0.012	2.456	3.237	-0.263	17.997	3.488	-0.012	**2.440**	3.142	-0.358	25.473

Loss*: The sum of squared errors for the fixed effect coefficients

Appendix B

This section includes the additional figures and tables in Chapter 3.

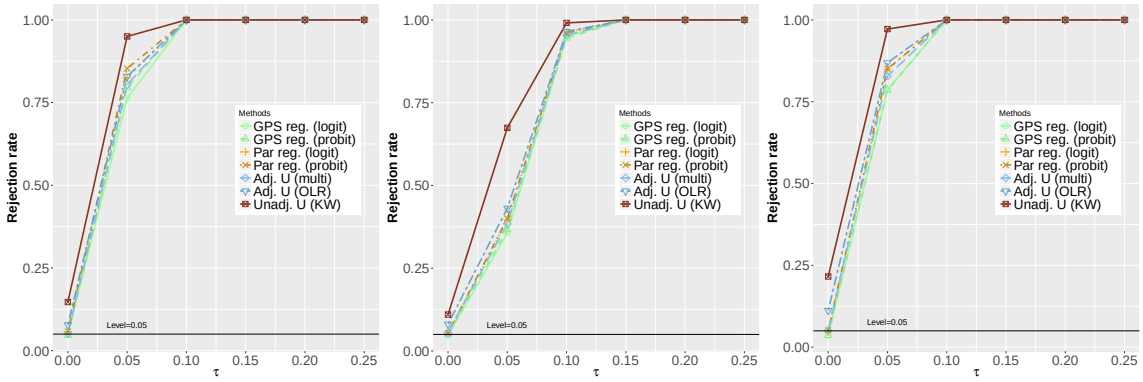
A.4 Other simulation results for ordinal outcome

Treatment generated from multinomial regression model

In this section, we reported the simulation results when the treatment was generated from the multinomial regression model. Similar to Section 3.4.2, the response was generated under three different regression models. The sizes of the overall tests under null hypothesis are given in Table A1.3 which shows that the actual size of the adjusted U-statistic is close to the nominal size when the GPS estimating model is correctly specified. We varied the treatment effect τ, and the power curves for different methods are reported in Figure A1.3. From Figure A1.3, it is clear that the adjusted U-statistics under correct specification of GPS has similar power curves as the parametric models. We also plotted the biases for estimating the superiority scores in Figures A1.4-A1.6 when response variable was generated from Boxcox, logit and probit models, respectively. The performances of the adjusted U-statistics are better than that of parametric models and regular U-statistics.

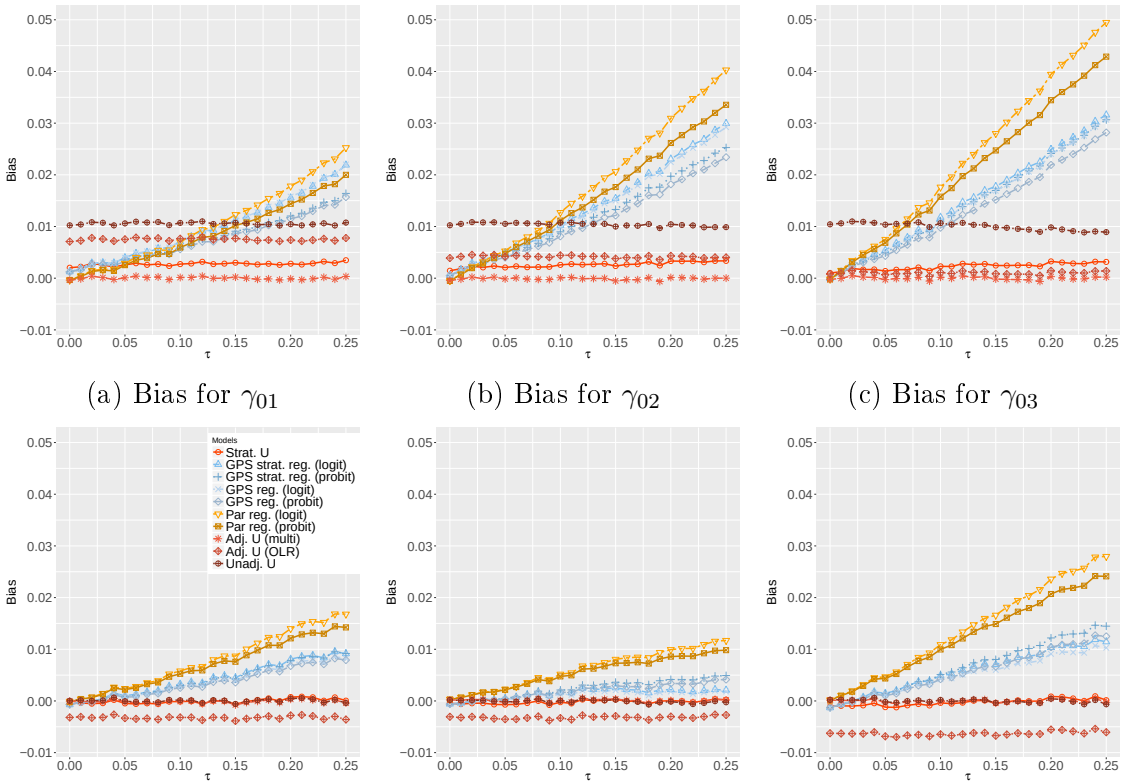
Table A1.3: Rejection rate for the overall test for different methods where the treatment assignment was generated from multinomial regression model, and outcome variable was generated from ordinal logit model (OR1), probit model (OR2), and Box-cox model (OR3) respectively

Methods	Outcome models		
	Logit	**Probit**	**Boxcox**
GPS regression (logit)	0.047	0.036	0.050
GPS regression (probit)	0.05	0.036	0.046
Parametric regression (logit)	0.054	0.047	0.058
Parametric regression (probit)	0.054	0.052	0.058
Adjusted U (Multinomial)	0.055	0.052	0.052
Adjusted U (OLR)	0.081	0.112	0.078
Unadjusted U (KW)	0.110	0.216	0.147



(a) Boxcox response model (b) Logit response model (c) Probit response model

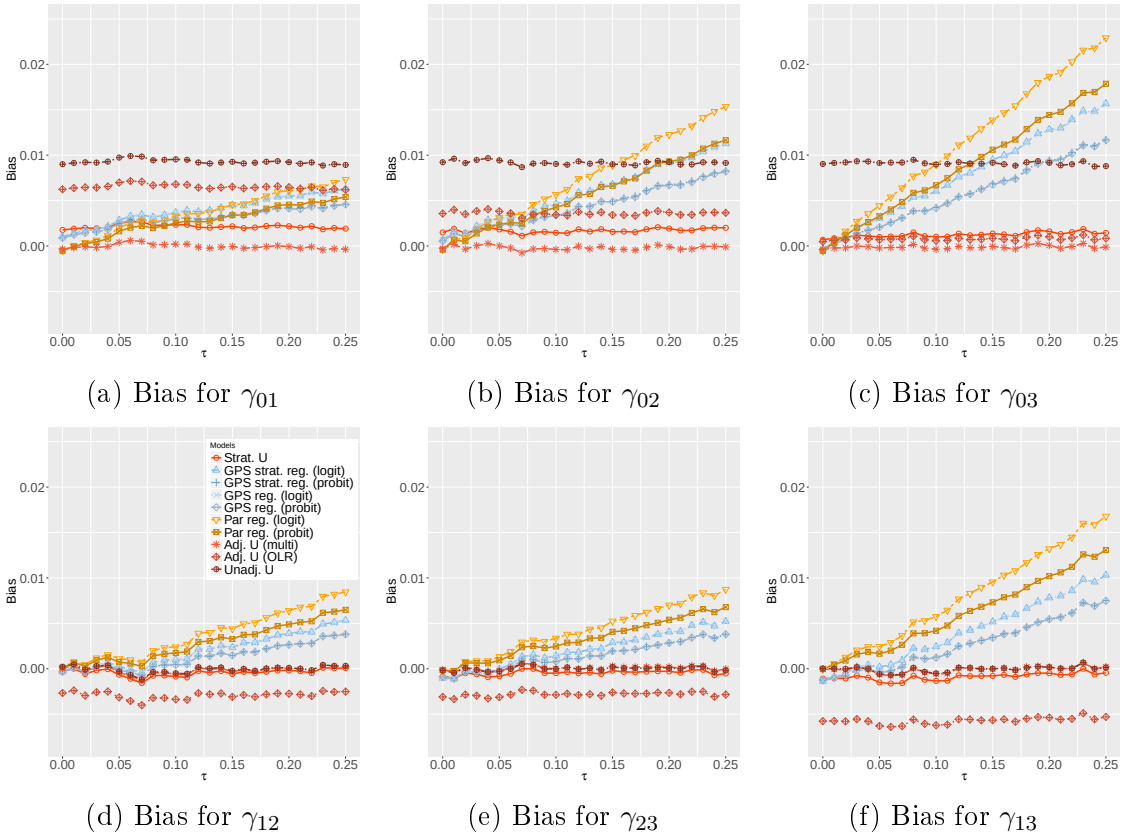
Figure A1.3: The power curves for testing overall effect for all methods when treatment was generated from multinomial regression



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

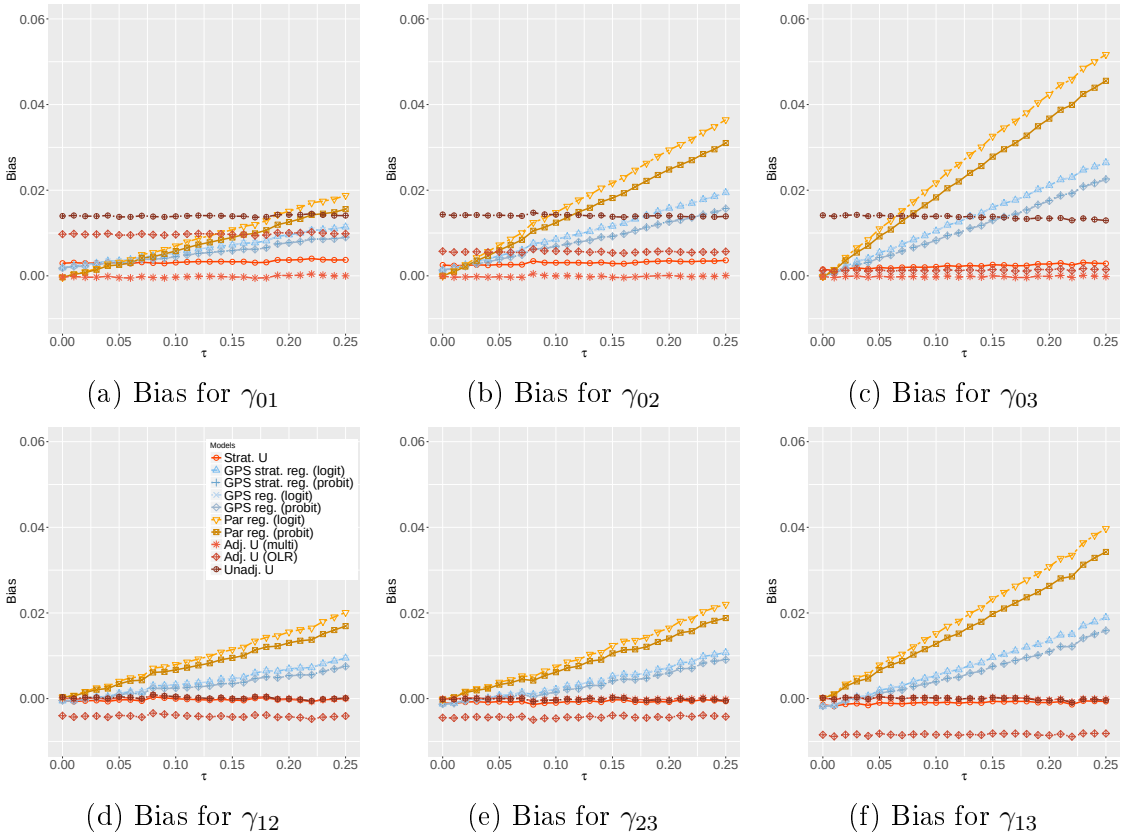
Figure A1.4: Bias plot for superiority scores when response was generated from Boxcox model and treatment was generated from multinomial logistic regression



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

Figure A1.5: Bias plot for superiority scores when response was generated from logit model and treatment was generated from multinomial logistic regression



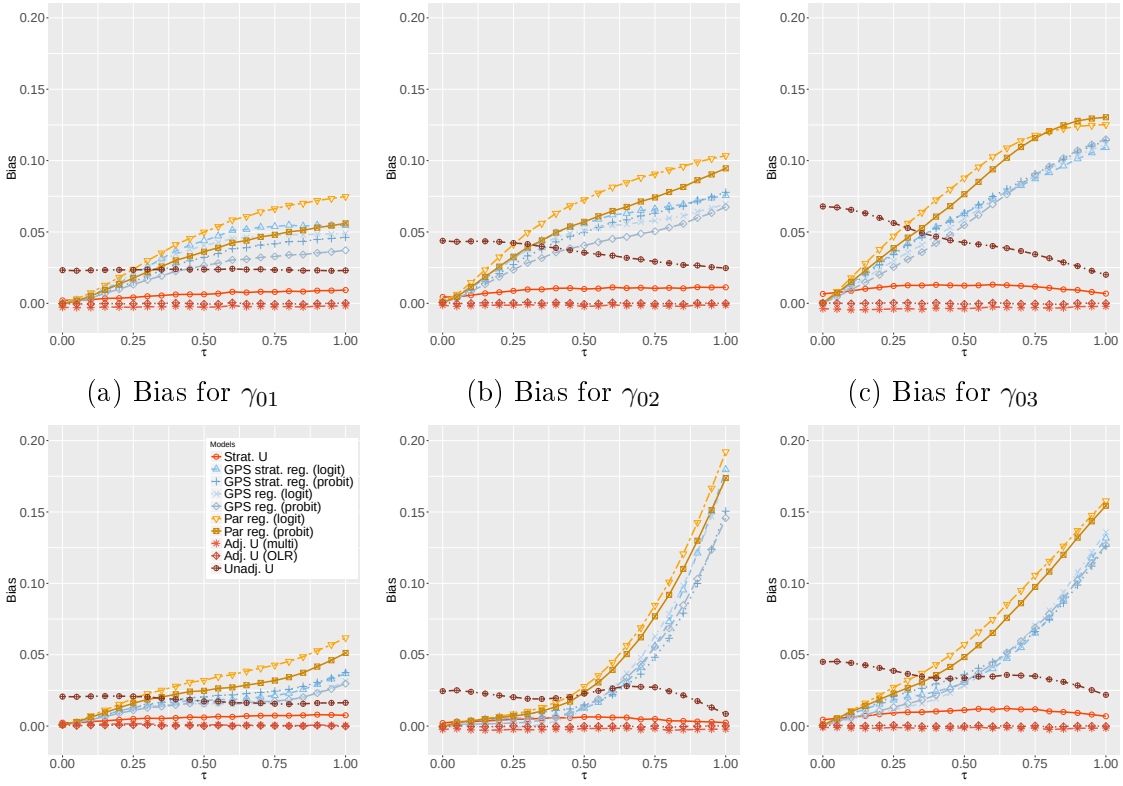
(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

Figure A1.6: Bias plot for superiority scores when response was generated from probit model and treatment was generated from multinomial regression

Simulation results when the correlation between confounding variables is high

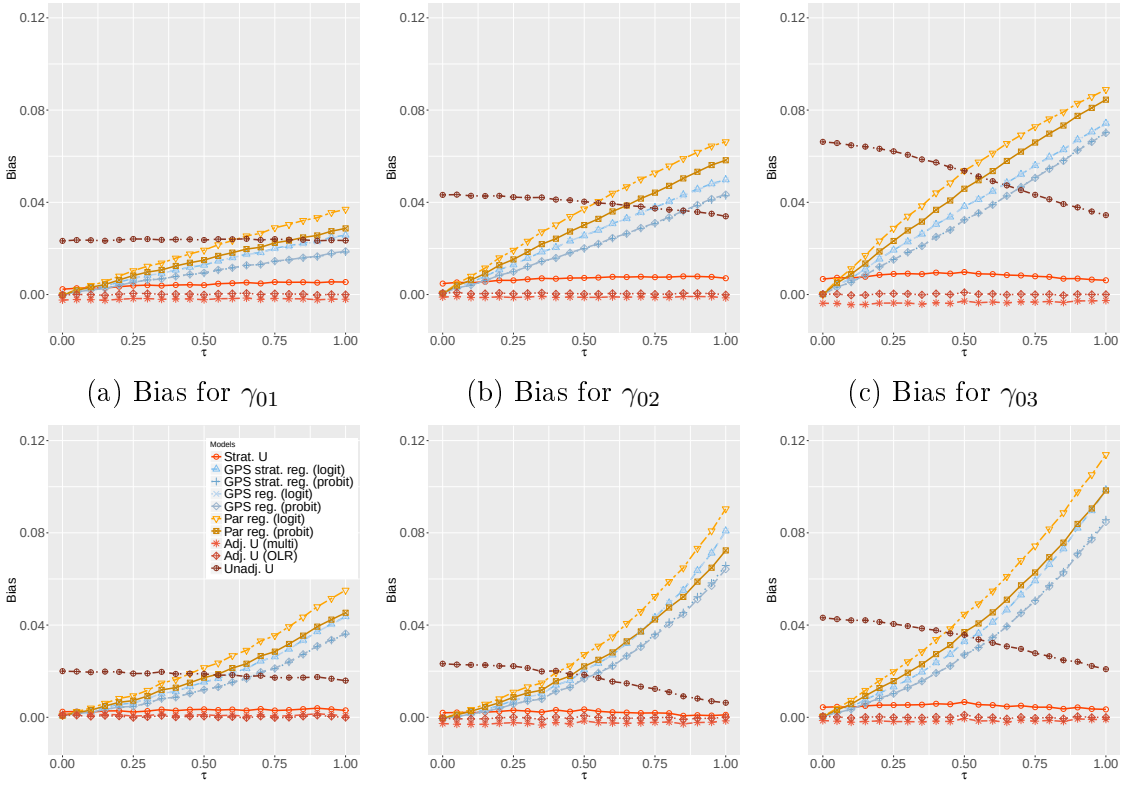
We also carried out simulations when the covariates are highly correlated. We have generated the same number of covariates as in Section 3.4.1 but with $(\rho_1, \rho_2) = (0.5, 0.6)$. Figures A1.7-A1.9 demonstrate the bias plots for estimation of all superiority scores when the the treatment was generated from ordinal logistic regression model and the response was generated respectively from the Box-cox, logit, and probit model. Figures A1.10-A1.12 show the bias plots for estimation of superiority scores when the treatment was generated from multinomial regression model and response was again generated under the three different models.



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

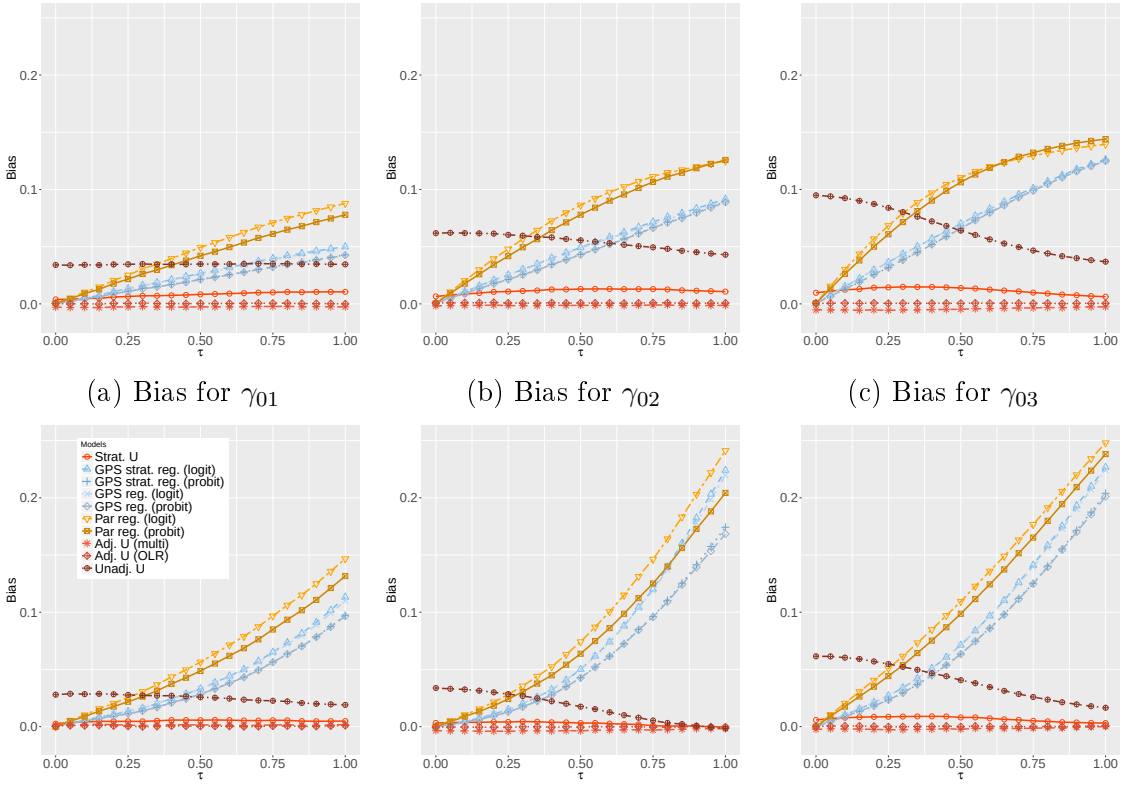
Figure A1.7: Bias plot for superiority scores when response was generated from Box-cox model, treatment was generated from ordinal, and confounding variables are highly correlated



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

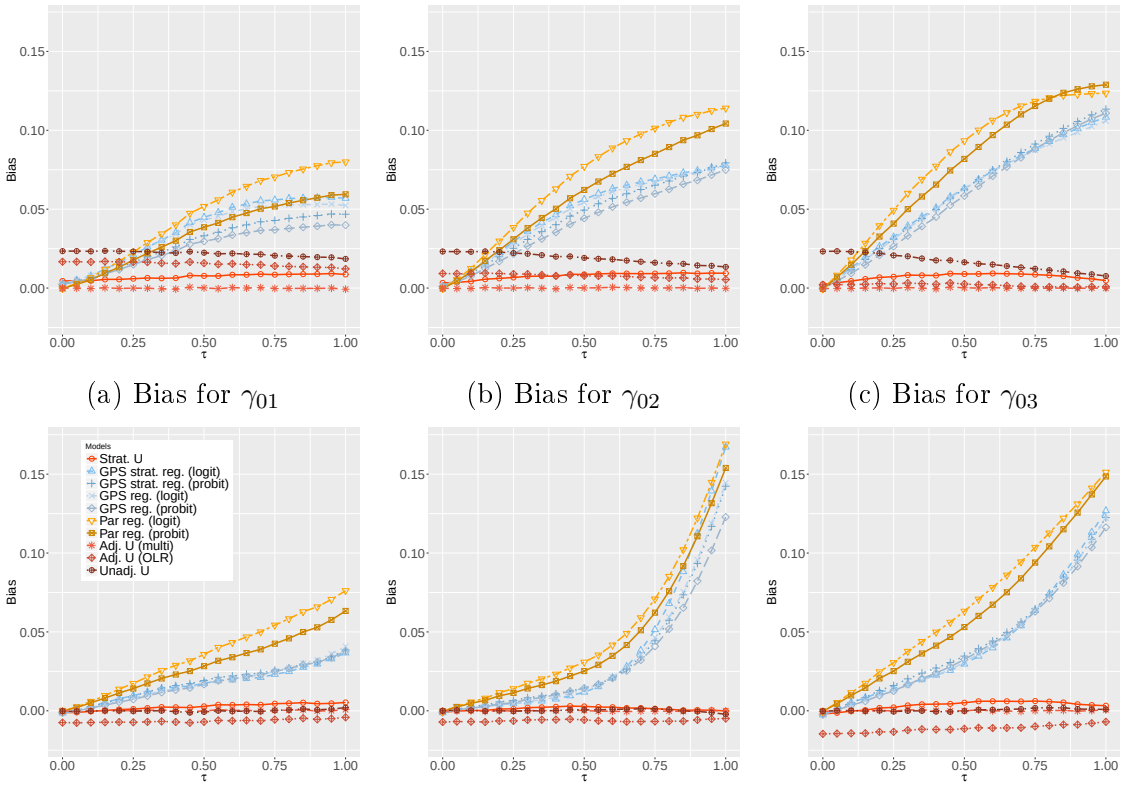
Figure A1.8: Bias plot for superiority scores when response was generated from ordinal logit model, treatment was generated from ordinal logit model, and confounding variables are highly correlated



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

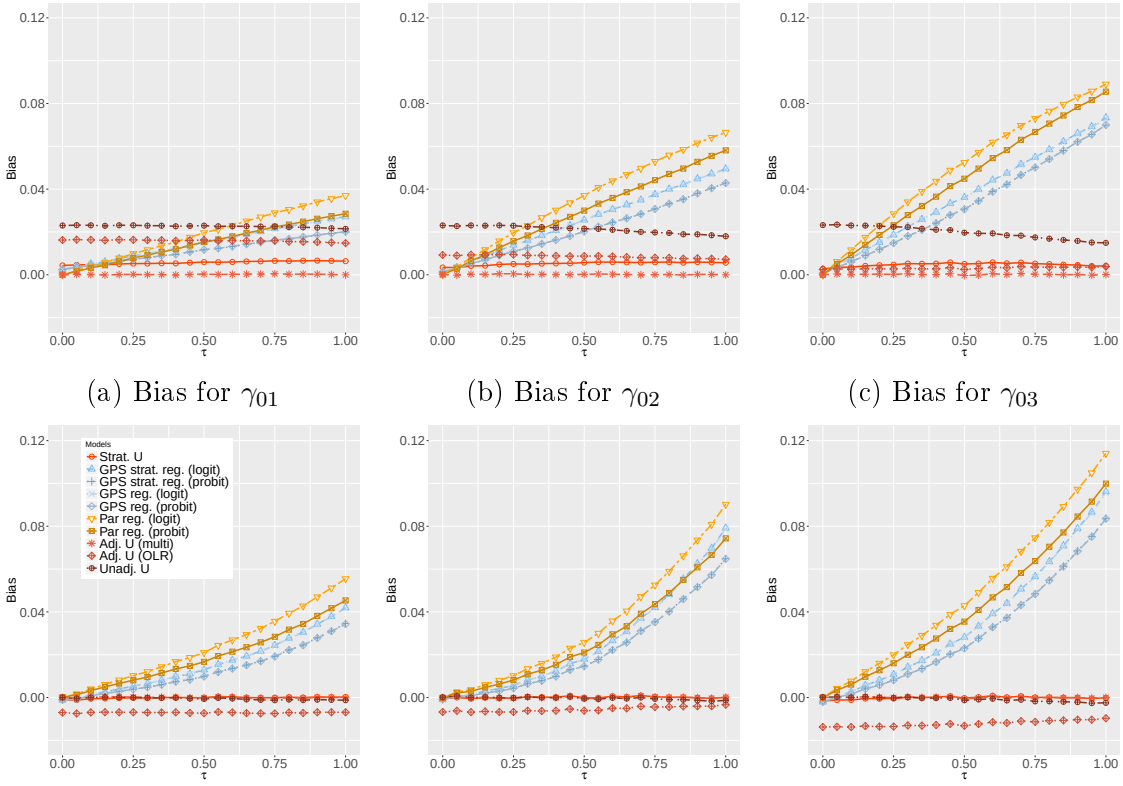
Figure A1.9: Bias plot for superiority scores when response was generated from probit model, treatment was generated from ordinal logit model, and confounding variables are highly correlated



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

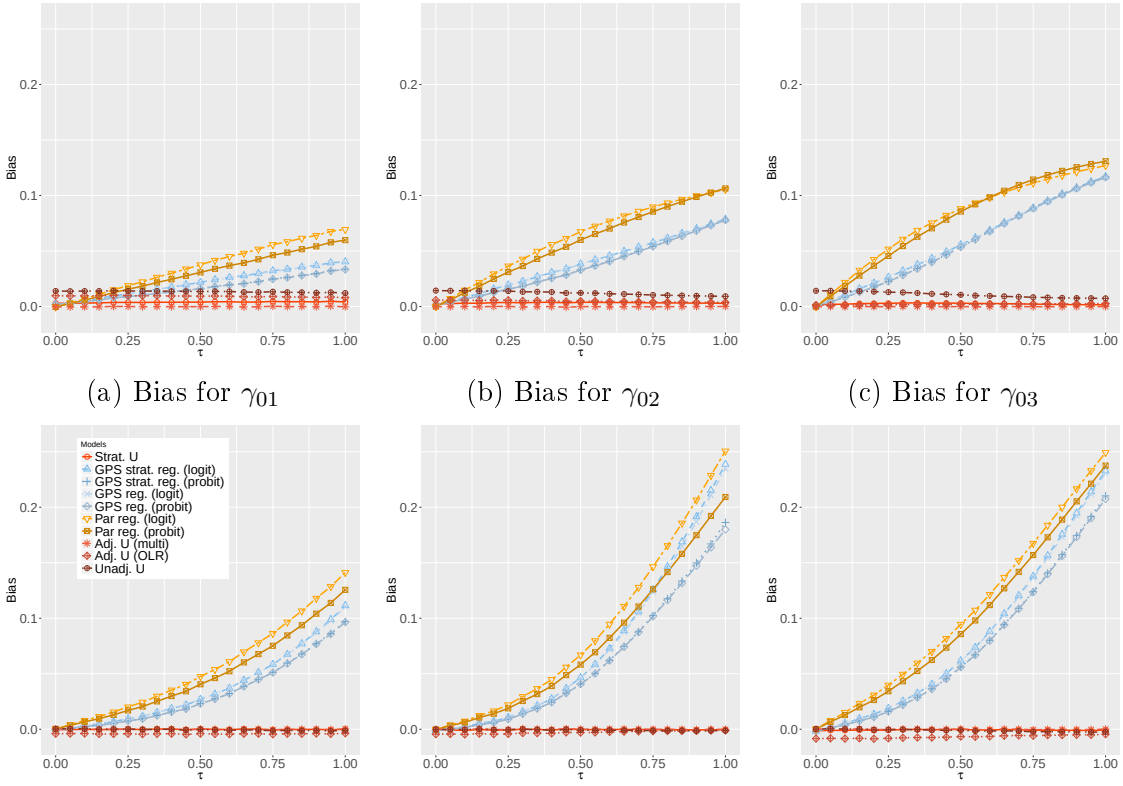
Figure A1.10: Bias plot for superiority scores when response was generated from Boxcox model, treatment was generated from multinomial logistic regression model, and confounding variables are highly correlated



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

Figure A1.11: Bias plot for superiority scores when response was generated from ordinal logistic regression model, treatment was generated from multinomial logistic regression model, and confounding variables are highly correlated



(a) Bias for γ_{01} (b) Bias for γ_{02} (c) Bias for γ_{03}

(d) Bias for γ_{12} (e) Bias for γ_{23} (f) Bias for γ_{13}

Figure A1.12: Bias plot for superiority scores when response was generated from probit model, treatment was generated from multinomial logistic regression model, and confounding variables are highly correlated

A.5 Score functions used in adjusted U-statistic

The score functions and the relevant derivatives used to construct adjusted U-statistics are presented here when the GPS is estimated using either multinomial regression or cumulative ordinal logistic regression respectively.

GPS estimated by multinomial regression model

When multinomial regression is used to estimate GPS, from

$$\log\left(\frac{P[T=k|X]}{P[T=1|X]}\right) = X\boldsymbol{\beta}^{(k)} \quad (k=1,\ldots,\mathcal{M}-1),$$

we obtain

$P[T=0|X] = \frac{1}{1+\sum_{k'=1}^{\mathcal{M}-1} e^{X\boldsymbol{\beta}^{(k')}}}$ and and $P[T=k|X] = p_{ik} = \frac{e^{X\boldsymbol{\beta}^{(k)}}}{1+\sum_{k'=1}^{\mathcal{M}-1} e^{X\boldsymbol{\beta}^{(k')}}}$ $(k=1,\ldots,\mathcal{M}-1)$.

Let denote t_{ik} as 1 if i^{th} subject receives treatment k. The log-likelihood for the i^{th} subject is given as

$$\begin{aligned}
l_i &= \sum_{k=0}^{\mathcal{M}-1} t_{ik}\log p_{ik} = -\sum_{k=0}^{\mathcal{M}-1} t_{ik}\left(\log\left(1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i\boldsymbol{\beta}^{(k')}}\right)\right) \\
&\quad + \sum_{k=1}^{\mathcal{M}-1} t_{ik}X_i\boldsymbol{\beta}^{(k)} \\
&= -\log\left(1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i\boldsymbol{\beta}^{(k')}}\right) + \sum_{k=1}^{\mathcal{M}-1} t_{ik}X_i\boldsymbol{\beta}^{(k)}, \text{ (since } \sum_{k=0}^{\mathcal{M}-1} t_{ik}=1). \\
\frac{\partial l_i}{\partial \boldsymbol{\beta}^{(k)}} &= -\frac{e^{X_i\boldsymbol{\beta}^{(k)}}}{1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i\boldsymbol{\beta}^{(k')}}}X_i + I_{\{t_i=k\}}X_i \\
&= \left(-p_{ik}+I_{\{t_i=k\}}\right)X_i, \text{where } k=1,\ldots,\mathcal{M}-1.
\end{aligned}$$

Thus, we have

$$\boldsymbol{S}_i = \frac{\partial l_i}{\partial \boldsymbol{\beta}} = (-p_{i1} + I_{\{t_i=1\}}, \ldots, -p_{i\mathcal{M}-1} + I_{\{t_i=\mathcal{M}-1\}}) \otimes X_i. \tag{4.33}$$

In addition, we also need to obtain

$$\begin{aligned}
\theta_i &= \frac{\partial \log w_i}{\partial \boldsymbol{\beta}} = -\frac{\partial \log p_{it_i}(x_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= \begin{cases} \dfrac{\partial \log(1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i' \boldsymbol{\beta}^{(k')}})}{\partial \boldsymbol{\beta}}, & \text{if } t_i = 0, \\ \dfrac{\partial[\log(1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i' \boldsymbol{\beta}^{(k')}}) - X_i \boldsymbol{\beta}^{(t_i)}]}{\partial \boldsymbol{\beta}}, & \text{if } t_i = 1, \ldots, \mathcal{M}-1, \end{cases} \\
&= \begin{cases} \dfrac{1}{1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i \boldsymbol{\beta}^{(k')}}} \begin{pmatrix} e^{X_i \boldsymbol{\beta}^{(1)}} X_i \\ \vdots \\ e^{X_i \boldsymbol{\beta}^{(\mathcal{M}-1)}} X_i \end{pmatrix}, & \text{if } t_i = 0, \\ \dfrac{\begin{pmatrix} e^{X_i \boldsymbol{\beta}^{(1)}} X_i \\ \vdots \\ e^{X_i \boldsymbol{\beta}^{(\mathcal{M}-1)}} X_i \end{pmatrix}}{1+\sum_{k'=1}^{\mathcal{M}-1} e^{X_i \boldsymbol{\beta}^{(k')}}} - \begin{pmatrix} I_{\{t_i=1\}} X_i \\ I_{\{t_i=2\}} X_i \\ \vdots \\ I_{\{t_i=\mathcal{M}-1\}} X_i \end{pmatrix}, & \text{if } t_i = 1, \ldots, \mathcal{M}-1, \end{cases} \\
&= \begin{pmatrix} p_{i1} - I_{\{t_i=1\}} \\ p_{i2} - I_{\{t_i=2\}} \\ \vdots \\ p_{i\mathcal{M}-1} - I_{\{t_i=\mathcal{M}-1\}} \end{pmatrix} \otimes X_i.
\end{aligned}$$

GPS estimated by ordinal logistic regression model

When the GPS is estimated from the ordinal logistic regression (OLR):

$$\log \frac{P[T \leq t|X}{1 - P[T \leq t|X]} = \alpha_t + X\beta \quad (t = 0, 1, \ldots, \mathcal{M} - 2), \tag{4.34}$$

the parameter in the GPS model can be written as

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{\mathcal{M}-2} \\ \beta \end{pmatrix}. \tag{4.35}$$

From equation (4.34), we have

$$P[T \leq t|X] = \frac{e^{\alpha_t + X\beta}}{1 + e^{\alpha_t + X\beta}}. \tag{4.36}$$

Thus,

$$\begin{aligned} P[T = 0|X] &= \frac{e^{\alpha_0 + X\beta}}{1 + e^{\alpha_0 + X\beta}}, \\ P[T = t|X] = P[T \leq t|X] - P[T \leq t - 1|X] &= \frac{e^{\alpha_t + X\beta}}{1 + e^{\alpha_t + X\beta}} - \frac{e^{\alpha_{t-1} + X\beta}}{1 + e^{\alpha_{t-1} + X\beta}}, \\ &\quad (t = 1, \ldots, \mathcal{M} - 2), \\ \text{and } P[T = \mathcal{M} - 1|X] &= \frac{1}{1 + e^{\alpha_{\mathcal{M}-2} + X\beta}}. \end{aligned}$$

Note that the log-likelihood for the entire sample is given by

$$
\begin{aligned}
& l(\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{M}-1}, \beta | X, T) \\
& = \sum_{i=1}^{N} l_i(\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{M}-1}, \beta | X_i, T_i) \\
& = \sum_{i=1}^{N} (t_{i0} \log p_{i0} + t_{i1} \log p_{i1} + \ldots + t_{i\mathcal{M}-1} \log p_{i\mathcal{M}-1}), \qquad (4.37)
\end{aligned}
$$

where $t_{it} = 1$ if the subject i receives k^{th} treatment, and $t_{it} = 0$ otherwise. Let denote $p_{it} = P[T_i = t | X = X_i]$. From (4.37) we obtain,

$$
\begin{aligned}
\frac{\partial l}{\partial \alpha_0} &= \sum_{i=1}^{N} \left(t_{i0} \frac{1}{p_{i0}} \frac{\partial p_{i0}}{\partial \alpha_0} + t_{i1} \frac{1}{p_{i1}} \frac{\partial p_{i1}}{\partial \alpha_0} \right) \\
&= \sum_{i=1}^{N} \left(\frac{t_{i0}}{p_{i0}} \frac{e^{\alpha_0 + X_i\beta}}{(1 + e^{\alpha_0 + X_i\beta})^2} - \frac{t_{i1}}{p_{i1}} \frac{e^{\alpha_0 + X_i\beta}}{(1 + e^{\alpha_0 + X_i\beta})^2} \right) \\
&= \sum_{i=1}^{N} \left(\frac{t_{i0}}{p_{i0}} - \frac{t_{i1}}{p_{i1}} \right) \frac{e^{\alpha_0 + X_i\beta}}{(1 + e^{\alpha_0 + X_i\beta})^2}, \\
\frac{\partial l}{\partial \alpha_1} &= \sum_{i=1}^{N} \left(t_{i1} \frac{1}{p_{i1}} \frac{\partial p_{i1}}{\partial \alpha_1} + t_{i2} \frac{1}{p_{i2}} \frac{\partial p_{i2}}{\partial \alpha_1} \right) \\
&= \sum_{i=1}^{N} \left(\frac{t_{i1}}{p_{i1}} \frac{e^{\alpha_1 + X_i\beta}}{(1 + e^{\alpha_1 + X_i\beta})^2} - \frac{t_{i2}}{p_{i2}} \frac{e^{\alpha_1 + X_i\beta}}{(1 + e^{\alpha_1 + X_i\beta})^2} \right) \\
&= \sum_{i=1}^{N} \left(\frac{t_{i1}}{p_{i1}} - \frac{t_{i2}}{p_{i2}} \right) \frac{e^{\alpha_1 + X_i\beta}}{(1 + e^{\alpha_1 + X_i\beta})^2}.
\end{aligned}
$$

Similarly,

$$
\frac{\partial l}{\partial \alpha_t} = \sum_{i=1}^{N} \left(\frac{t_{it}}{p_{it}} - \frac{t_{it+1}}{p_{it+1}} \right) \frac{e^{\alpha_t + X_i\beta}}{(1 + e^{\alpha_t + X_i\beta})^2}, \quad (t = 2, \ldots, \mathcal{M} - 2).
$$

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= \sum_{i=1}^{N} \left(\sum_{k=0}^{\mathcal{M}-1} \frac{t_{ik}}{p_{ik}} \frac{\partial p_{ik}}{\partial \beta} \right) \\
&= \sum_{i=1}^{N} \left[\frac{t_{i0}}{p_{i0}} \cdot \frac{e^{\alpha_0 + X_i\beta}}{(1+e^{\alpha_0+X_i\beta})^2} + \sum_{t=1}^{\mathcal{M}-2} \frac{t_{it}}{p_{it}} \left(\frac{e^{\alpha_t+X_i\beta}}{(1+e^{\alpha_t+X_i\beta})^2} - \frac{e^{\alpha_{t-1}+X_i\beta}}{(1+e^{\alpha_{t-1}+X_i\beta})^2} \right) \right. \\
&\quad \left. - \frac{t_{i\mathcal{M}-1}}{p_{i\mathcal{M}-1}} \cdot \frac{e^{\alpha_{\mathcal{M}-2}+X_i\beta}}{(1+e^{\alpha_{\mathcal{M}-2}+X_i\beta})^2} \right] X_i \\
&= \sum_{i=1}^{N} \left[\sum_{t=0}^{\mathcal{M}-2} \left(\frac{t_{it}}{p_{it}} - \frac{t_{it+1}}{p_{it+1}} \right) \frac{e^{\alpha_t+X_i\beta}}{(1+e^{\alpha_t+X_i\beta})^2} \right] X_i.
\end{aligned}$$

Denote $p_{it} = P[T = t|X = X_i]$ and $w_i(X_i; \boldsymbol{\beta}) = \frac{1}{P[t_i=t|X_i,\boldsymbol{\beta}]} = \frac{1}{p_{it_i}}$. Note that,

$$p_{it} = \begin{cases} p_{i0} = \dfrac{e^{\alpha_0+X_i\beta}}{1+e^{\alpha_0+X_i\beta}}, & \text{if } t = 0; \\ \dfrac{e^{\alpha_t+X_i\beta}}{1+e^{\alpha_t+X_i\beta}} - \dfrac{e^{\alpha_{t-1}+X_i\beta}}{1+e^{\alpha_{t-1}+X_i\beta}}, & \text{if } t = 1, \ldots, \mathcal{M}-2; \\ \dfrac{1}{1+e^{\alpha_{\mathcal{M}-2}+X\beta}}, & \text{if } t = \mathcal{M}-1. \end{cases}$$

Let us denote $p^c_{it} = P[T \leq t|X_i]$, then

$$\frac{\partial p^c_{it}}{\partial \boldsymbol{\beta}} = \frac{e^{\alpha_t+X\beta}}{(1+e^{\alpha_t+X\beta})^2} \cdot \frac{\partial (\alpha_t + X\beta)}{\partial \boldsymbol{\beta}}.$$

Let E_t denote a vector of length $\mathcal{M}-1$ with all zero except the t^{th} element as 1. Then we have,

$$\frac{\partial \log w_i(X_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\partial \log p_{it_i}}{\partial \boldsymbol{\beta}} = -\frac{1}{p_{it_i}}\frac{\partial p_{it_i}}{\partial \boldsymbol{\beta}}$$

$$= \begin{cases} -\dfrac{1}{p_{i0}} p_{i0} \begin{pmatrix} E_1 \\ X_i \end{pmatrix}, \\ \text{if } t_i = 0; \\ -\dfrac{1}{p_{it_i}} \left[p^c_{it_i}\left(1-p_{it_i}\right) \begin{pmatrix} E_{t_i+1} \\ X_i \end{pmatrix} - p^c_{it_{i-1}}\left(1-p_{it_{i-1}}\right) \begin{pmatrix} E_{t_i} \\ X_i \end{pmatrix} \right], \\ \text{if } t_i = 1,\ldots,\mathcal{M}-2; \\ -\dfrac{1}{p_{i\mathcal{M}-1}} \left[-p^c_{i\mathcal{M}-2}\left(1-p_{i\mathcal{M}-2}\right) \begin{pmatrix} E_{\mathcal{M}-1} \\ X_i \end{pmatrix} \right], \\ \text{if } t_i = \mathcal{M}-1. \end{cases}$$

$$\therefore \frac{\partial \log w_i(X_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{cases} \left(p_{i0}-1\right) \begin{pmatrix} E_1 \\ X_i \end{pmatrix}, & \text{if } t_i = 0; \\ \begin{pmatrix} \frac{p^c_{it_{i-1}}\left(1-p^c_{it_{i-1}}\right)}{p_{it_i}} E_{t_i} + \frac{p^c_{it_i}\left(p^c_{it_i}-1\right)}{p_{it_i}} E_{t_{i+1}} \\ \left(p^c_{it_i} + p^c_{it_{i-1}} - 1\right) X_i \end{pmatrix}, & \text{if } t_i = 1,\ldots,\mathcal{M}-2; \\ p^c_{i\mathcal{M}-2} \begin{pmatrix} E_{\mathcal{M}-1} \\ X_i \end{pmatrix}, & \text{if } t_i = \mathcal{M}-1. \end{cases}$$

Appendix C

This section includes the building blocks of the R-codes used in the three projects.

A.6 R code and STAN code for Hierarchical Mixed Effect Hurdle Model for Time and Spatially Correlated Count Data and its Bayesian Analysis

The STAN code and the R code of the important parts of Chapter 2 are provided in this section.

STAN code:

```
 ### Negative Binomial Hurdle Model
### New Full Data, m=26 variables
### County Spline, fixed phi
data {
int <lower=1> N;
int <lower=1> m;
int <lower=1> k;
int <lower=1> T;
int <lower=1> st_n;
int <lower=1> stc_n;
int <lower=1> P;
int <lower=0> y[N];
vector[T-1] mu0;
matrix[N, m] X_fixed;
matrix[N, P] X_statetime;
matrix[N, T-1] X_time;
matrix[N, k] X_spline;
matrix[stc_n, k] prelim_mat;
cov_matrix[T-1] V1;
```

```
cov_matrix[T-1] V2;
real  d;
real nu;
cov_matrix[T] Sigma;
real  phi;
}
transformed data {
vector[T] mu;
for (i in 1:T) mu[i] = 0;
}
parameters{
vector[m] gamma;
vector[m] beta;
real  a;
vector[k] sp;
vector[P] sT;
real <lower=0> sigma_sp;
cov_matrix[T] psi;
vector[T-1] time1;
vector[T-1] time2;
}
model{
for(n in 1:N){
 if (y[n] == 0)
target += binomial_logit_lpmf(0|1,X_fixed[n,]*beta+
a*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time1);
else
target += binomial_logit_lpmf(1|1,X_fixed[n,]*beta+
a*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time1)
```

```
+ neg_binomial_2_log_lpmf(y[n] | X_fixed[n,]*gamma+
d*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time2,phi)-
neg_binomial_2_lccdf(0 | exp(X_fixed[n,]*gamma+
d*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time2),phi);
}
for(j in 1:m){
gamma[j]~normal(0,10);
beta[j]~normal(0,10);
}
a~normal(0,10);
#phi~gamma(1,1);
for(i in 0:(st_n-1))
{
sT[(T*i+1):(T*(i+1))] ~ multi_normal(mu,psi);
}
psi ~ inv_wishart(nu, Sigma);

for(i in 1:k){
sp[i] ~ normal(0,sigma_sp);
}
sigma_sp ~ cauchy(0,0.1)T[0,];
time1~multi_normal(mu0,V1);
time2~multi_normal(mu0,V2);
}
generated quantities{
vector[N] loglik_store;
vector[stc_n] check;
real loglik_total;
```

```
for(n in 1:N){
if (y[n] == 0)
loglik_store[n] = binomial_logit_lpmf(0|1,X_fixed[n,]*beta+
a*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time1);
else
loglik_store[n] = binomial_logit_lpmf(1|1,X_fixed[n,]*beta+
a*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time1)
+ neg_binomial_2_log_lpmf(y[n] | X_fixed[n,]*gamma+
d*(X_statetime[n,]*sT+X_spline[n,]*sp)+
X_time[n,]*time2,phi)
- neg_binomial_2_lccdf(0 | exp(X_fixed[n,]*gamma+
d*(X_statetime[n,]*sT+X_spline[n,]*sp)+X_time[n,]*time2),phi);
}
for(j in 1:stc_n){
check[j] = prelim_mat[j,]*sp;
}
loglik_total = sum(loglik_store);
}
```

R code:

```
rm(list=ls())
args <- commandArgs()
## start w/args[3]
Mod <- as.numeric(args[3])
Sim <- as.numeric(args[4])
fname <- paste("NBfull␣phi␣0.12", Mod, Sim, "RData", sep = ".")
setwd("/home/s0ghos07/NegBin")
data.new <- read.table(file="FINAL.txt")
```

```
library(mvtnorm)
library(MASS)
library(geosphere)
library(plyr)
library(maps)
library(mapproj)
library(hexbin)
library(ggplot2)
library(httr)
library(dplyr)
library(stringr)
library(rstan)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

########## making the knot points
subd <- subset(data.new,select=c("lat","long","state","county"))
subd <- unique(subd)
coord <- subd[,1:2]
### making knots
lat20 <- round(seq(min(coord[,1]),max(coord[,1]),
by = ((max(coord[,1])-min(coord[,1]) )/(15 - 1))),4)
long15 <- round(seq(min(coord[,2]),max(coord[,2]),
by = ((max(coord[,2])-min(coord[,2]) )/(20 - 1))),4)
knot.mat <- arrange(expand.grid(lat20 =lat20 ,long15 =long15),
lat20 )
knot.mat <- as.data.frame(knot.mat)
names(knot.mat) <- c("Lat","Long")
prelim.mat1 <- matrix(rep(0,dim(coord)[1]*dim(knot.mat)[1]),
```

```
byrow=TRUE,ncol=dim(knot.mat)[1])
dist.knot <- function (long1, lat1, long2, lat2)
#### calculating knots
{
  rad <- pi/180
  a1 <- lat1 * rad
  a2 <- long1 * rad
  b1 <- lat2 * rad
  b2 <- long2 * rad
  dlon <- b2 - a2
  dlat <- b1 - a1
  a <- (sin(dlat/2))^2 + cos(a1) * cos(b1) * (sin(dlon/2))^2
  c <- 2 * atan2(sqrt(a), sqrt(1 - a))
  R <- 6378.145
  d <- round(0.621371*R * c,2)
  #knot <- exp(-d)
  return(d)
}
for(i in 1: dim(coord)[1]){
  for(j in 1: dim(knot.mat)[1]){
    prelim.mat1[i,j] <- round(dist.knot(coord[i,2],coord[i,1],
    knot.mat[j,2],knot.mat[j,1]),4)
  }
}
matrix <- knot.mat[which(apply(prelim.mat1,2,min)<100),]
knot.mat <- matrix ### OUR DESIRED NUMBER OF KNOTS

################################################
####### subsetting the data for ##############
```

```
####### fitting GLM to get initial values ####
################################################

data.new <- data.new[,-c(39,40,41)]
colnames(data.new)
subset <- data.new[,c(1,7:40)]
subset1 <- data.new[,c(7:22,24,26:29,34,37,39,40)]

##########################################
######## Defining some constants #########
##########################################

st.n <- length(unique(data.new$state))  ### no. of states
stc.n <- length(unique(data.new$county)) ### no. of counties
T <- length(unique(data.new$year)) ### no. of years
mu0 <- as.vector(rep(0,T))
var1 <- var2 <- diag(T)
N <- dim(data.new)[1] ### total data in medium data
V1 <- V2 <- 100*diag(T-1)
mu0 <- rep(0,T-1)
nu <- 7
Sigma <- (1/T)*diag(T)
P <- T*st.n
Var <- diag(st.n)
m <- 1+dim(subset1)[2]   ### no. of covariates
############################
######### Indexing #########
############################
####### Renaming the factors of state into numbers
```

```
index.state <- data.new$state
for(i in 1:length(levels(data.new$state))){
  levels(index.state)[levels(index.state)==
  levels(data.new$state)[i]] <- i
}
index.state <- factor(index.state)
#### indexing for county
index.county <- data.new$county
for(i in 1:length(levels(data.new$county))){
levels(index.county)[levels(index.county)==
levels(data.new$county)[i]] <- i
}
#### indexing for year
data.new$year <- factor(data.new$year)
index.year <- data.new$year
for(i in 1:length(levels(data.new$year))){
  levels(index.year)[levels(index.year)==
  levels(data.new$year)[i]] <- i
}
#### indexing for statetime
sttm.2 <- paste(index.state,index.year,sep=".")
sttm.2 <- factor(sttm.2)
index.statetime <- sttm.2
for(i in 1:length(levels(sttm.2))){
levels(index.statetime)[levels(index.statetime)==
levels(sttm.2)[i]] <- i
}
#length(index.statetime)
############################
```

```
#### design matrices #######
############################
X.fixed <- round(as.matrix(data.frame(rep(1,N),
subset1[,1:dim(subset1)[2]])),3)
X.time <- matrix(rep(0,T*N),ncol=T)
for(i in 1:T){
  X.time[,i] <- ifelse(index.year==i,1,0)
}
X.time <- X.time[,1:5]
#sum(apply(X.time,2,sum))

X.state <- matrix(rep(0,st.n*N),ncol=st.n)
for(i in 1:st.n){
  X.state[,i] <- ifelse(index.state==i,1,0)
}
X.county <- matrix(rep(0,stc.n*N),ncol=stc.n)
for(i in 1:stc.n){
  X.county[,i] <- ifelse(index.county==i,1,0)
}
X.statetime <- matrix(rep(0,N*P),ncol=P)
for(i in 1:P){
  X.statetime[,i] <- ifelse(index.statetime==i,1,0)
}
################################################
######## Computation of Distance Matrix #########
################################################
library(geosphere)
sbst <- cbind(data.new$county, data.new$long, data.new$lat)
sbst <- unique(sbst)
```

```
coord <- sbst[,-1]
prelim.mat <- matrix(rep(0,dim(coord)[1]*dim(knot.mat)[1]),
byrow=TRUE,ncol=dim(knot.mat)[1])
for(i in 1: dim(coord)[1]){
  for(j in 1: dim(knot.mat)[1]){
    prelim.mat[i,j] <- round(dist.knot(coord[i,1],coord[i,2],
    knot.mat[j,2],knot.mat[j,1]),4)
  }
}
prelim.mat <- (1/50)*prelim.mat
prelim.mat <- round(exp(-prelim.mat),3)
spline.design <- prelim.mat[rep(seq_len(nrow(prelim.mat)),
 each=T),]
################################################
###### Create Initial values from above ######
################################################
norandom <- read.table("random_negbin_1.txt")
# norandom is the stan output without random effect
gamma.init <- norandom[1:m,1]
beta.init <- norandom[(m+1):(2*m),1]
gamma.var <- matrix(rep(0,m*m),ncol=m,byrow = TRUE)
beta.var <- matrix(rep(0,m*m),ncol=m,byrow = TRUE)
diag(gamma.var) <- norandom[1:m,2]
diag(beta.var) <- norandom[(m+1):(2*m),2]
gamma.table <- matrix(rep(0,m*3),nrow=m,ncol=3)
beta.table <- matrix(rep(0,m*3),nrow=m,ncol=3)
s.table <- matrix(rep(0,P*3),nrow=P,ncol=3)
time1.table <- matrix(rep(0,(T-1)*3),nrow=T-1,ncol=3)
time2.table <- matrix(rep(0,(T-1)*3),nrow=T-1,ncol=3)
```

```
a.table <- b.table <- d.table <- sigma.sp.table <-
sigma_county.table <- e.table <- inv_rho_sq.table <- rep(0,3)
psi.table <- array(data=NA,dim = c(T,T,3))
sp.table <- matrix(rep(0,dim(knot.mat)[1]*3),
nrow=dim(knot.mat)[1],ncol=3)
setseed(10000)
gamma.table[,1] <- round(mvrnorm(1,gamma.init,gamma.var),3)
beta.table[,1] <- round(mvrnorm(1,beta.init,beta.var),3)
#s.table[,1] <- round(as.numeric(mvrnorm(1,rep(0,st.n),
0.3*diag(st.n))),3)
time1.table[,1] <- round(as.numeric(mvrnorm(1,rep(0,5),
0.1*diag(5))),3)
time1.table[,1] <- round(as.numeric(mvrnorm(1,rep(0,5),
0.11*diag(5))),3)
a.table[1] <- round(rnorm(1,1,0.1),3)
b.table[1] <- round(rnorm(1,1,0.1),3)
d.table[1] <- round(rnorm(1,1,0.1),3)
e.table[1] <- round(rnorm(1,1,0.1),3)
psi.table[,,1] <- diag(T)
sp.table[,1] <- round(rnorm(dim(knot.mat)[1],0,0.01),3)
sigma.sp.table[1] <- round(abs(rcauchy(1,0,0.1)),3)
set.seed(10001)
gamma.table[,2] <- round(mvrnorm(1,gamma.init,gamma.var),3)
beta.table[,2] <- round(mvrnorm(1,beta.init,beta.var),3)
#s.table[,2] <- round(as.numeric(mvrnorm(1,rep(0,st.n),
0.3*diag(st.n))),3)
time1.table[,2] <- round(as.numeric(mvrnorm(1,rep(0,5),
0.1*diag(5))),3)
time1.table[,2] <- round(as.numeric(mvrnorm(1,rep(0,5),
```

```
0.11*diag(5))),3)
a.table[2] <- round(rnorm(1,1,0.1),3)
b.table[2] <- round(rnorm(1,1,0.1),3)
d.table[2] <- round(rnorm(1,1,0.1),3)
e.table[2] <- round(rnorm(1,1,0.1),3)
psi.table[,,2] <- diag(T)
sp.table[,2] <- round(rnorm(dim(knot.mat)[1],0,0.01),3)
sigma.sp.table[2] <- round(abs(rcauchy(1,0,0.1)),3)
set.seed(10002)
gamma.table[,3] <- round(mvrnorm(1,gamma.init,gamma.var),3)
beta.table[,3] <- round(mvrnorm(1,beta.init,beta.var),3)
#s.table[,3] <- round(as.numeric(mvrnorm(1,rep(0,st.n),
0.3*diag(st.n))),3)
time1.table[,3] <- round(as.numeric(mvrnorm(1,rep(0,5),
0.1*diag(5))),3)
time1.table[,3] <- round(as.numeric(mvrnorm(1,rep(0,5),
0.11*diag(5))),3)
a.table[3] <- round(rnorm(1,1,0.1),3)
b.table[3] <- round(rnorm(1,1,0.1),3)
d.table[3] <- round(rnorm(1,1,0.1),3)
e.table[3] <- round(rnorm(1,1,0.1),3)
psi.table[,,3] <- diag(T)
sp.table[,3] <- round(rnorm(dim(knot.mat)[1],0,0.01),3)
sigma.sp.table[3] <- round(abs(rcauchy(1,0,0.1)),3)
#####################################################
######## Defining some constants again  #########
#####################################################

st.n <- length(unique(data.new$state))  ### no. of states
```

```
stc.n <- length(unique(data.new$county)) ### no. of counties
T <- length(unique(data.new$year)) ### no. of years
mu0 <- as.vector(rep(0,T))
var1 <- var2 <- diag(T)
N <- dim(data.new)[1] ### total data in medium data
V1 <- V2 <- 100*diag(T-1)
mu0 <- rep(0,T-1)
nu <- 7
Sigma <- (1/T)*diag(T)
P <- T*st.n
Var <- diag(st.n)
m <- 1+dim(subset1)[2]   ### no. of covariates
k <- dim(knot.mat)[1]   ###  no. of knot points

#####################################################
######### Added state, county effect #############
#####################################################

full_negbin_data <- list(N=N,m=m,k=k, T=T, st_n=st.n,
stc_n=stc.n, mu0=mu0, y=data.new$y, X_fixed=X.fixed,
P=P, nu=nu, Sigma=Sigma,  X_spline=spline.design,
prelim_mat=prelim.mat, X_statetime=X.statetime,
X_time=X.time,  V1=V1, V2=V2, d=1, phi=1/0.12)
init_list <- list(list(gamma=gamma.table[,1],
beta=beta.table[,1],sT=s.table[,1],a= a.table[1],
psi =psi.table[,,1],time1=time1.table[,1],
time2=time2.table[,1],sp=sp.table[,1],
sigma_sp=sigma.sp.table[1]), list(gamma=gamma.table[,2],
beta=beta.table[,2],sT=s.table[,2], a= a.table[2],
```

```
psi =psi.table[,,2], time1=time1.table[,2],
time2=time2.table[,2],sp=sp.table[,2],
sigma_sp=sigma.sp.table[1]), list(gamma=gamma.table[,3],
beta=beta.table[,3],sT=s.table[,3], a= a.table[3],
psi =psi.table[,,3], time1=time1.table[,3],
time2=time2.table[,3],sp=sp.table[,3],
sigma_sp=sigma.sp.table[3]))

# 3 chains
ptm <- proc.time()
full_fit <- stan(file = 'negbin_statetime_spline.stan',
data = full_negbin_data, control=list(metric="diag_e"),
init = init_list, iter = 600, warmup=300, chains = 3,
seed= 11627)
proc.time()-ptm
save(full_fit,file=fname)
```

A.7 R code for Comparisons of Average Treatment Effects for Multiple Groups when Outcome is Ordinal and Confounding Exists

The important functions used for generating data, simulation and case study in Chapter 3 are provided below.

Functions:

```
BoxCox <- function(x, lambda){
          Fx <- sapply(x,function(x) {
          if(lambda==0) stop("lambda should not be zero")
          if(lambda>0 & x < -1/lambda){return(0)}
          if(lambda<0 & x > -1/lambda){return(1)}
          if(lambda!=0 & (1+lambda*x)>=0){
```

```
                return((1+lambda*x)^(1/lambda)/((1+lambda*x)^
                (1/lambda)+1))
          }
          })
          return(Fx=Fx)
}
BoxCox_Inverse <- function(q,lambda){
          ((q/(1-q))^lambda-1)/lambda
}

#### Generating the data
Generate.Data <- function(N,tau,gamma.true,Beta,Treat.Model,
                          Resp.Model,seed)
     {
      set.seed(seed)

      ########################################
      ## Generate covariates
      ########################################
         rho1=0.1
         rho2=0.2
         mu=rep(0,6)
         stddev=rep(1,6)
        corMat1 <- diag(length(mu))
        corMat1[outer(1:nrow(corMat1), 1:ncol(corMat1),
        function(i,j) i!=j)] <- rho1
        corMat2 <- diag(length(mu))
        corMat2[outer(1:nrow(corMat2), 1:ncol(corMat2),
        function(i,j) i!=j)] <- rho2
```

```
  covMat1 <- stddev %*% t(stddev) * corMat1
  covMat2 <- stddev %*% t(stddev) * corMat2
  dat1 <- mvrnorm(n = N, mu = mu, Sigma = covMat1,
  empirical = FALSE)
  dat2 <- mvrnorm(n = N, mu = mu, Sigma = covMat2,
  empirical = FALSE)

X1=dat1[,1];X2=dat1[,2];X3=dat1[,3];X4=dat1[,4];
X5=dat1[,5];X6=dat1[,6];
X7=dat2[,1];X8=dat2[,2];X9=dat2[,3];X10=dat2[,4];
X11=dat2[,5];X12=dat2[,6];
z <- data.frame(rep(1,N),X1,X2,X3,X4,X5,X6,X7,X8,
X9,X10,X11,X12)
colnames(z) <- NULL
z <- as.matrix(z)
  z.trt <- z[,c(1:4,8:10)]
  z.resp <- z[,1:7]
##########################################
## Generate Group Assignment
##########################################
  if(Treat.Model=="Multinomial")
  {
  Beta1 = Beta[,1]
  Beta2 = Beta[,2]
  Beta3 = Beta[,3]
  Beta4 = Beta[,4]
TProb = cbind(exp(z.trt%*%Beta1), exp(z.trt%*%Beta2),
 exp(z.trt%*%Beta3), exp(z.trt%*%Beta4))
  sum <- apply(TProb, 1, sum)
```

```
        TProb <- TProb/sum
        TChoices = t(apply(TProb, 1, rmultinom, n = 1, size = 1))
        g <- apply(TChoices, 1, function(x)
        which(x==1))
        Trt <- TChoices[,2:4]
        ng <- dim(Trt)[2]
        }
        if(Treat.Model=="OLR")
        {
        ## Drawing probabilities from OLR
        alph <- c(-log(3),0,log(3))
 #beta.true <- c(0,-0.1,-0.2,-0.3,-0.3,-0.2,-0.1)
 z1 <- exp(alph[1]-z.trt%*%beta.true)
 z2 <- exp(alph[2]-z.trt%*%beta.true)
 z3 <- exp(alph[3]-z.trt%*%beta.true)
 p1 <- z1/(1+z1)
 p2 <- z2/(1+z2)-p1
 p3 <- z3/(1+z3)-p1-p2
 p4 <- 1-p1-p2-p3
TProb <- cbind(p1,p2,p3,p4)

      # multinomial draws
        TChoices = t(apply(TProb, 1, rmultinom, n = 1, size= 1))
        g <- apply(TChoices, 1, function(x) which(x==1))
        Trt <- TChoices[,2:4]
        #ng <- dim(Trt)[2]+1
        #table(g)
        }
      ##########################################
```

```
   ## Generate Response
   ########################################

   if (Resp.Model=="Probit"){
      alpha <- c(qnorm(0.25),0,qnorm(0.75))

### Control
z1 <- alpha[1]-z.resp%*%gamma.true
z2 <- alpha[2]-z.resp%*%gamma.true
z3 <- alpha[3]-z.resp%*%gamma.true
p1 <- pnorm(z1)
p2 <- pnorm(z2)-p1
p3 <- pnorm(z3)-p1-p2
p4 <- 1-p1-p2-p3

 p <- cbind(p1,p2,p3,p4)
 temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
 Y0 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

        ### T1
       z1 <- alpha[1]-tau[1]-z.resp%*%gamma.true
       z2 <- alpha[2]-tau[1]-z.resp%*%gamma.true
       z3 <- alpha[3]-tau[1]-z.resp%*%gamma.true
       p1 <- pnorm(z1)
       p2 <- pnorm(z2)-p1
       p3 <- pnorm(z3)-p1-p2
       p4 <- 1-p1-p2-p3

       p <- cbind(p1,p2,p3,p4)
```

```
temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
Y1 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

  ### T2
z1 <- alpha[1]-tau[2]-z.resp%*%gamma.true
z2 <- alpha[2]-tau[2]-z.resp%*%gamma.true
z3 <- alpha[3]-tau[2]-z.resp%*%gamma.true
p1 <- pnorm(z1)
p2 <- pnorm(z2)-p1
p3 <- pnorm(z3)-p1-p2
p4 <- 1-p1-p2-p3

p <- cbind(p1,p2,p3,p4)
temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
Y2 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

  ### T3
z1 <- alpha[1]-tau[3]-z.resp%*%gamma.true
z2 <- alpha[2]-tau[3]-z.resp%*%gamma.true
z3 <- alpha[3]-tau[3]-z.resp%*%gamma.true
p1 <- pnorm(z1)
p2 <- pnorm(z2)-p1
p3 <- pnorm(z3)-p1-p2
p4 <- 1-p1-p2-p3

p <- cbind(p1,p2,p3,p4)
temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
Y3 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N
Y <- ifelse(g==1,Y0,ifelse(g==2,Y1,ifelse(g==3,Y2,Y3)))
```

```
        }

if (Resp.Model=="Logit"){
          alpha <- c(-log(3),0,log(3))

            ### Control
           z1 <- exp(alpha[1]-z.resp%*%gamma.true)
           z2 <- exp(alpha[2]-z.resp%*%gamma.true)
           z3 <- exp(alpha[3]-z.resp%*%gamma.true)
           p1 <- z1/(1+z1)
           p2 <- z2/(1+z2)-p1
           p3 <- z3/(1+z3)-p1-p2
           p4 <- 1-p1-p2-p3

           p <- cbind(p1,p2,p3,p4)
           temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
           Y0 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

            ### T1
           z1 <- exp(alpha[1]-tau[1]-z.resp%*%gamma.true)
           z2 <- exp(alpha[2]-tau[1]-z.resp%*%gamma.true)
           z3 <- exp(alpha[3]-tau[1]-z.resp%*%gamma.true)
           p1 <- z1/(1+z1)
           p2 <- z2/(1+z2)-p1
           p3 <- z3/(1+z3)-p1-p2
           p4 <- 1-p1-p2-p3

           p <- cbind(p1,p2,p3,p4)
           temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
```

```
Y1 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

  ### T2
z1 <- exp(alpha[1]-tau[2]-z.resp%*%gamma.true)
z2 <- exp(alpha[2]-tau[2]-z.resp%*%gamma.true)
z3 <- exp(alpha[3]-tau[2]-z.resp%*%gamma.true)
p1 <- z1/(1+z1)
p2 <- z2/(1+z2)-p1
p3 <- z3/(1+z3)-p1-p2
p4 <- 1-p1-p2-p3

p <- cbind(p1,p2,p3,p4)
temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
Y2 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

  ### T3
z1 <- exp(alpha[1]-tau[3]-z.resp%*%gamma.true)
z2 <- exp(alpha[2]-tau[3]-z.resp%*%gamma.true)
z3 <- exp(alpha[3]-tau[3]-z.resp%*%gamma.true)
p1 <- z1/(1+z1)
p2 <- z2/(1+z2)-p1
p3 <- z3/(1+z3)-p1-p2
p4 <- 1-p1-p2-p3

p <- cbind(p1,p2,p3,p4)
temp.y <- apply(p,1,function(x) rmultinom(1,1,x))
Y3 <- apply(temp.y,2,which.max)  #;Y;table(Y)/N

Y <- ifelse(g==1,Y0,ifelse(g==2,Y1,ifelse(g==3,Y2,Y3)))
```

```
}
if (Resp.Model=="Box-Cox"){
    alpha1 <- BoxCox_Inverse(0.25,1)
    alpha2 <- BoxCox_Inverse(0.5,1)
    alpha3 <- BoxCox_Inverse(0.75,1)
    alpha <- c(alpha1,alpha2,alpha3)

      ###### Potential Outcomes ######

    ## F1 ##
    lambda <- 1+tau[1]

    z11 <- alpha[1]-tau[1]-z.resp%*%gamma.true
    z12 <- alpha[2]-tau[1]-z.resp%*%gamma.true
    z13 <- alpha[3]-tau[1]-z.resp%*%gamma.true
    p11 <- BoxCox(z11,lambda)
    p12 <- BoxCox(z12,lambda)-p11
    p13 <- BoxCox(z13,lambda)-p11-p12
    p14 <- 1-p11-p12-p13

    p1 <- cbind(p11,p12,p13,p14)
    temp.y1 <- apply(p1,1,function(x) rmultinom(1,1,x))
    Y1 <- apply(temp.y1,2,which.max)  #;Y;table(Y)/N

    ## F2 ##
    lambda <- 1+tau[2]

    z11 <- alpha[1]-tau[2]-z.resp%*%gamma.true
    z12 <- alpha[2]-tau[2]-z.resp%*%gamma.true
```

```
z13 <- alpha[3]-tau[2]-z.resp%*%gamma.true
p11 <- BoxCox(z11,lambda)
p12 <- BoxCox(z12,lambda)-p11
p13 <- BoxCox(z13,lambda)-p11-p12
p14 <- 1-p11-p12-p13

p2 <- cbind(p11,p12,p13,p14)
temp.y2 <- apply(p2,1,function(x) rmultinom(1,1,x))
Y2 <- apply(temp.y2,2,which.max)  #;Y;table(Y)/N

## F3 ##
lambda <- 1+tau[3]

z11 <- alpha[1]-tau[3]-z.resp%*%gamma.true
z12 <- alpha[2]-tau[3]-z.resp%*%gamma.true
z13 <- alpha[3]-tau[3]-z.resp%*%gamma.true
p11 <- BoxCox(z11,lambda)
p12 <- BoxCox(z12,lambda)-p11
p13 <- BoxCox(z13,lambda)-p11-p12
p14 <- 1-p11-p12-p13

p3 <- cbind(p11,p12,p13,p14)
temp.y3 <- apply(p3,1,function(x) rmultinom(1,1,x))
Y3 <- apply(temp.y3,2,which.max)  #;Y;table(Y)/N

## F0 ##
z01 <- alpha[1]-z.resp%*%gamma.true
z02 <- alpha[2]-z.resp%*%gamma.true
z03 <- alpha[3]-z.resp%*%gamma.true
```

```
            p01 <- BoxCox(z01,1)
            p02 <- BoxCox(z02,1)-p01
            p03 <- BoxCox(z03,1)-p01-p02
            p04 <- 1-p01-p02-p03
              p0 <- cbind(p01,p02,p03,p04)
              temp.y0 <- apply(p0,1,function(x) rmultinom(1,1,x))
            Y0 <- apply(temp.y0,2,which.max)  #;Y;table(Y)/N

            Y <- ifelse(g==1,Y0,ifelse(g==2,Y1,ifelse(g==3,Y2,Y3)))
            #Y <- as.factor(Y)
        }


        return(data=list(y=Y,Trt=Trt,g=g,z=z,Y1=Y1,Y2=Y2,
        Y3=Y3,Y0=Y0))

}

### Wilcoxon kernal functions

wilcoxon = function(y0,y1) {
    mat.kernel=mat.or.vec(length(y0), length(y1))
    wil.knl<-function(x0,x1){(x0<x1)+0.5*(x0==x1)}
    mat.kernel<-outer(y0, y1, wil.knl)
    return(mat.kernel)
    }

##########################################################
```

```
### Function to calculate the multiple-sample adj U-stat
#############################################################

#############################################################
###### Adjusted U stat (GPS from Multinomial) #########
#############################################################

ssus.multi.ho=function(data,kernel.name='wilcoxon')
  { kernel.function=match.fun(kernel.name)
    n.tot=length(data$g)  # no. of obs, group assignment in g
    groups<-sort(unique(data$g))
    ng=length(unique(data$g))   # how many groups
    npg<-as.vector(table(data$g)) #no of obs per group
    nv=dim(as.matrix(data$z))[2]  # no of covariates
    wt<-prop.score<-rep(0,n.tot)
    np=(ng-1)*nv # number of parameters in multinomial reg
    s=as.matrix( mat.or.vec(n.tot, np) )  # si in the
        #estimating equation,
    # the matrix the derivative of loglikelihood
    #function gamma
    theta=as.matrix(mat.or.vec(n.tot, np))
    ## the matrix: the derivative of ln(w(zi, gi, gamma))
    ## w.r.t. gamma
    ## note that the order for gamma is based on each
    ## fixed covariate.
    c.n=rep(0,np)
    prop.score.model=vglm(data$g~data$z+0,
    family=multinomial(refLevel=1))
        # summary(prop.score.model)
```

```
    fv=fitted.values(prop.score.model) #n*np
    jinv=-vcov(prop.score.model)
    ## the inverse of the second derivative w.r.t. gamma
    ## Calculate Si in the estimating equation for
    ## estimating gamma

  ##  calculate weights
    for (i in 1:n.tot)
    {prop.score[i]=fv[i,data$g[i]]  #pr[G=gi|xi]
     wt[i]<-1/prop.score[i]   #wt[i]=1/p[G=gi|x])
     }
     fv0=-fv
     for (i in 1:n.tot) {fv0[i,data$g[i]]=fv0[i,data$g[i]]+1}
     fv1<-fv0[,c(-1)]
     for (i in 1:n.tot) {s[i,]=as.vector(fv1[i,]%o%data$z[i,])}

    ##  Define theta, the derivative of ln w(zi,gi,gamma)
    ##wrt gamma using KW
     for (i in 1:n.tot) {
    group.index<-rep(0, ng); group.index[data$g[i]]<-1
    temp<-fv[i,c(-1)]-group.index[c(-1)]
    theta[i,]<- as.vector(temp%o%as.vector(data$z[i,]))
     }

#############################################################
##  The following comparing every other versus group 1
#############################################################
  SET1<-SET2<-U<-SD<-P.Val<-c()  #Install all parwise
  ## comparisons
```

```
u.hat<-rep(0,ng-1)
var.u<-mat.or.vec(ng-1,ng-1)
psai<-matrix(NA, nrow=n.tot, ncol=ng-1)
set1<-c(1)
#j<-2
for (j in 2:ng)
{ set2<-c(j)
ind1<-which(data$g%in%set1)
ind2<-which(data$g%in%set2)
n1<-length(ind1); n2<-length(ind2)
w1<-mean(wt[ind1])
w2<-mean(wt[ind2])
kernel.mat<-wilcoxon(data$y[ind1], data$y[ind2])
u.hat[j-1]<-u.hat1<-t(wt[ind1]/w1)%*%kernel.mat%*%
(wt[ind2]/w2)/(n1*n2)
 ################################################
#  Calculate variance
################################################
temp1<-(wt[ind1]/w1)%o%rep(1,length(ind2))
temp2<-rep(1,length(ind1))%o%(wt[ind2]/w2)
kernel.wt<-temp1*kernel.mat*temp2

## Define Cn ##
Cn <- rep(-Inf, np)
A<-theta[ind1,]; B<-theta[ind2,]
#p<-1
 for (p in 1:np){
 C.temp <- outer(A[,p],B[,p],"+")
 CK <- kernel.wt*C.temp
```

```
    Cn[p] <- u.hat1*mean(wt[ind1]*A[,p])+
    u.hat1*mean(wt[ind2]*B[,p])-mean(CK)
    }
    Cn.3<-t(Cn)%*%jinv%*%t(s)/n.tot     #the third term
    ## for pasai
    psi1=(-u.hat1*as.vector(t(wt[ind1]/w1)-1)+
    as.vector(apply(kernel.wt, 1,mean)-u.hat1))/n1+Cn.3[ind1]
    psi2=(-u.hat1*as.vector(t(wt[ind2]/w2)-1)+
    as.vector(apply(kernel.wt, 2,mean)-u.hat1))/n2+Cn.3[ind2]
    psai[ind1,j-1]<-psi1; psai[ind2,j-1]<-psi2
    set12<-union(set1, set2)
    var.u[j-1,j-1]<-var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2
    SET1<-c(SET1,set1)
    SET2<-c(SET2,set2)
    U<-c(U,u.hat1)
    SD<-c(SD,sqrt(var.12))
    P.Val<-c(P.Val, 2*pnorm(-abs(u.hat[j-1]-0.5)/
    sqrt(var.12)))
       }
 #### Perform the overall test
for (i in 2:(ng-1))       # u-stat i vs 1
  { for (j in (i+1):ng)       # u-stat j vs 1
     {
      ind0<-which(data$g==1); ind1<-which(data$g==i);
      ind2<-which(data$g==j)
      n0<-length(ind0); n1<-length(ind1);
      n2<-length(ind2);
      temp<-(n0^2/((n0-1)^2))*sum(outer(psai[ind0,(i-1)],
      psai[ind0,(j-1)], "*"))
```

```
      #temp<-sum(psai[ind0,(i-1)]*psai[ind0,(j-1)])
      temp<-temp+(n0*n2)/((n0-1)*(n2-1))*sum(outer(psai[ind0,
      (i-1)],psai[ind2,(j-1)], "*"))
      temp<-temp+(n1*n0)/((n1-1)*(n0-1))*sum(outer(psai[ind1
      ,(i-1)],psai[ind0,(j-1)], "*"))
      temp<-temp+(n1*n2)/((n1-1)*(n2-1))*sum(outer(psai[ind1
      ,(i-1)],psai[ind2,(j-1)], "*"))
      var.u[i-1,j-1]<-var.u[j-1,i-1]<-temp
       }
         }
      Stat.Wald<-t(u.hat-0.5)%*%solve(var.u)%*%(u.hat-0.5)
      p.val<-pchisq(Stat.Wald, df=ng-1, ncp = 0,
      lower.tail = FALSE)
####### Pairwise comparison
for (i in 2:(ng-1))
  { set1<-c(i)
   for (j in (i+1):ng)
    { set2<-c(j)
      ind1<-which(data$g%in%set1)
      ind2<-which(data$g%in%set2)
      n1<-length(ind1); n2<-length(ind2)
      w1<-mean(wt[ind1])
      w2<-mean(wt[ind2])
      kernel.mat<-wilcoxon(data$y[ind1],
      data$y[ind2])
      u.hat12<-t(wt[ind1]/w1)%*%kernel.mat%*%
      (wt[ind2]/w2)/(n1*n2)
      ################################################
      #  Calculate variance
```

```
  ################################################
  temp1<-(wt[ind1]/w1)%o%rep(1,length(ind2))
  temp2<-rep(1,length(ind1))%o%(wt[ind2]/w2)
  kernel.wt<-temp1*kernel.mat*temp2
  tild.h1<- apply(kernel.wt,1,mean)
  tild.h2<- apply(kernel.wt,2,mean)
  ## Define Cn ##
  Cn <- rep(-Inf, np)
  A<-theta[ind1,]; B<-theta[ind2,]
  #p<-1
 for (p in 1:np){
  C.temp <- outer(A[,p],B[,p],"+")
  CK <- kernel.wt*C.temp
  Cn[p] <- u.hat12*mean(wt[ind1]*A[,p])+
  u.hat12*mean(wt[ind2]*B[,p])-mean(CK)
       }
Cn.3<-t(Cn)%*%jinv%*%t(s)/n.tot     #the third term
## for pasai
psi1=(-u.hat12*as.vector(t(wt[ind1]/w1)-1)+
as.vector(apply(kernel.wt, 1,mean)-u.hat12))/n1+Cn.3[ind1]
psi2=(-u.hat12*as.vector(t(wt[ind2]/w2)-1)+
as.vector(apply(kernel.wt, 2,mean)-u.hat12))/n2+Cn.3[ind2]
#var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2+
##2*n1*n2*mean(outer(psi1, psi2, "*"))
    var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2
SET1<-c(SET1,set1)
SET2<-c(SET2,set2)
U<-c(U,u.hat12)
SD<-c(SD,sqrt(var.u[j-1, j-1]))
```

```
      P.Val<-c(P.Val, 2*pnorm(-abs(u.hat12-0.5)/sqrt(var.12)))
         }
       }
       Pairwise<-cbind(SET1, SET2, U, SD, P.Val)
       colnames(Pairwise)<-c("SET1", "SET2", "U-stat", "SD",
       "P.Val")
       out=list(U.hat =u.hat, Wald.Stat=Stat.Wald, P.val=p.val,
       Pairwise.Comparisons=Pairwise )
       return(out)
 }
################################################
###### Adjusted U stat (GPS from OLR) #########
################################################

Adjust.U.OLR=function(data,kernel.name='wilcoxon')
  { kernel.function=match.fun(kernel.name)
    n.tot=length(data$g)  # no of obs, group assignment in g
    groups<-sort(unique(data$g))
    ng=length(unique(data$g))   # how many groups
    npg<-as.vector(table(data$g)) #no of obs per group
    nv=dim(as.matrix(data$z))[2]  # number of covariates
    wt<-wt.tilde<-prop.score<-rep(0,n.tot)

      np=(ng-1)*nv # number of parameters in multinomial reg
        s=as.matrix( mat.or.vec(n.tot, np) )  # si in the
        ## estimating equation,
        # the matrix the derivative of loglikelihood func
        ## gamma
        theta=as.matrix(mat.or.vec(n.tot, np))
```

```
## the matrix: the derivative of ln(w(zi, gi, gamma))
## w.r.t. gamma
## note that the order for gamma is based on each
## fixed covariate.
c.n=rep(0,np)

prop.score.model=clm(factor(data$g)~data$z[,c(-1)],
link="logit")
fit1<-fitted.values(prop.score.model)
##  calculate weights
wt<-1/ fit1 #wt[i]=1/p[G=gi|x])

#    summary(prop.score.model)
jinv=-vcov(prop.score.model)
## the inverse of the second derivative w.r.t. gamma
## Calculate Si in the estimating equation for
## estimating gamma
xbeta<-data$z[,c(-1)]%*%coef(prop.score.model)
[ng:(ng+nv-2)]
Xbeta.alph<-mat.or.vec(n.tot, ng-1)
for (g in 1:(ng-1))
 { Xbeta.alph[,g]<-rep(coef(prop.score.model)[g],n.tot)
 -xbeta}
Cumul.p<-plogis(Xbeta.alph)  #Cumulative distribution

fv<-mat.or.vec(n.tot, ng)
fv[,1]<- Cumul.p[,1]; fv[,ng]<-1- Cumul.p[,ng-1];
for (g in 2:(ng-1))
 {fv[,g]<- Cumul.p[,g]- Cumul.p[,g-1]}
```

```
    # prodictted probablities

    np=(ng-1)+(nv-1) # number of parameters in ordinal reg
    s<-theta<-mat.or.vec(n.tot, np)  # si derivetive
    ## li/gamma
    #i<-1
    for (i in 1:n.tot)
      { ci<-c(rep(0, ng-1),data$z[i,c(-1)])
        gi<-data$g[i]
        if(data$g[i]==1)
           {ci[gi]<-1; s[i,]<-(1-fv[i,1])*ci}
        if(gi>1 & gi<ng)
           {ci[gi-1]<-(-1)*Cumul.p[i,gi-1]*(1-Cumul.p[i,gi-1])
           /fit1[i]
            ci[gi]<-Cumul.p[i,gi]*(1-Cumul.p[i,gi])
            /fit1[i]
            ci[ng:np]<-(1-Cumul.p[i,gi-1]-Cumul.p[i,gi])
            *ci[ng:np]
            s[i,]<-ci
           }
        if(gi==ng)
           {ci[ng-1]<-1; s[i,]<-(-1)*Cumul.p[i,ng-1]*ci}
        theta[i,]<-(-1)*s[i,]
       }

##############################################################
##  The following comparing every other versus group 1
##############################################################
    SET1<-SET2<-U<-SD<-P.Val<-c()  #Install all parwise
```

```
## comparisons
u.hat<-rep(0,ng-1)
var.u<-mat.or.vec(ng-1,ng-1)
psai<-matrix(NA, nrow=n.tot, ncol=ng-1)
set1<-c(1)
#j<-2
for (j in 2:ng)
 { set2<-c(j)
   ind1<-which(data$g%in%set1)
   ind2<-which(data$g%in%set2)
   n1<-length(ind1); n2<-length(ind2)
   w1<-mean(wt[ind1])
   w2<-mean(wt[ind2])
   kernel.mat<-wilcoxon(data$y[ind1], data$y[ind2])
   u.hat[j-1]<-u.hat1<-t(wt[ind1]/w1)%*%kernel.mat%*%
   (wt[ind2]/w2)/(n1*n2)
   ################################################
   #  Calculate variance
   ################################################
   temp1<-(wt[ind1]/w1)%o%rep(1,length(ind2))
   temp2<-rep(1,length(ind1))%o%(wt[ind2]/w2)
   kernel.wt<-temp1*kernel.mat*temp2

   ## Define Cn ##
   Cn <- rep(-Inf, np)
   A<-theta[ind1,]; B<-theta[ind2,]
   #p<-1
   for (p in 1:np){
      C.temp <- outer(A[,p],B[,p],"+")
```

```
      CK <- kernel.wt*C.temp
      Cn[p] <- u.hat1*mean(wt[ind1]*A[,p])+
      u.hat1*mean(wt[ind2]*B[,p])-mean(CK)
    }

  Cn.3<-t(Cn)%*%jinv%*%t(s)/n.tot    #the third term
  ## for pasai
  psi1=(-u.hat1*as.vector(t(wt[ind1]/w1)-1)+
  as.vector(apply(kernel.wt, 1,mean)-u.hat1))/
  n1+Cn.3[ind1]
  psi2=(-u.hat1*as.vector(t(wt[ind2]/w2)-1)+
  as.vector(apply(kernel.wt, 2,mean)-u.hat1))/
  n2+Cn.3[ind2]
  psai[ind1,j-1]<-psi1; psai[ind2,j-1]<-psi2
  set12<-union(set1, set2)
  #for (g in set12)
  #  {ind<-which(data$g==g)
  #   psai[ind,j-1]<-psai[ind,j-1]-mean( psai[ind,j-1])
  #  }
  #var.u[j-1,j-1]<-var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2+
  2*n1*n2*mean(outer(psi1, psi2, "*"))
   var.u[j-1,j-1]<-var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2
  SET1<-c(SET1,set1)
  SET2<-c(SET2,set2)
  U<-c(U,u.hat1)
  SD<-c(SD,sqrt(var.12))
  P.Val<-c(P.Val, 2*pnorm(-abs(u.hat[j-1]-0.5)/
  sqrt(var.12)))
}
```

```
#### Perform the overall test
  for (i in 2:(ng-1))       # u-stat i vs 1
    { for (j in (i+1):ng)       # u-stat j vs 1
       {
        ind0<-which(data$g==1); ind1<-which(data$g==i);
        ind2<-which(data$g==j)
        n0<-length(ind0); n1<-length(ind1);
        n2<-length(ind2);
        temp<-(n0^2/((n0-1)^2))*sum(outer(psai[ind0,(i-1)],
        psai[ind0,(j-1)], "*"))
        #temp<-sum(psai[ind0,(i-1)]*psai[ind0,(j-1)])
        temp<-temp+(n0*n2)/((n0-1)*(n2-1))*sum(outer(
        psai[ind0,(i-1)],psai[ind2,(j-1)], "*"))
        temp<-temp+(n1*n0)/((n1-1)*(n0-1))*sum(
        outer(psai[ind1,(i-1)],psai[ind0,(j-1)], "*"))
        temp<-temp+(n1*n2)/((n1-1)*(n2-1))*sum(outer(
        psai[ind1,(i-1)],psai[ind2,(j-1)], "*"))
        var.u[i-1,j-1]<-var.u[j-1,i-1]<-temp
       }
     }
  Stat.Wald<-t(u.hat-0.5)%*%solve(var.u)%*%(u.hat-0.5)
  p.val<-pchisq(Stat.Wald, df=ng-1, ncp = 0,
  lower.tail = FALSE)
####### Pairwise comparison
for (i in 2:(ng-1))
  { set1<-c(i)
    for (j in (i+1):ng)
     { set2<-c(j)
```

```
  ind1<-which(data$g%in%set1)
  ind2<-which(data$g%in%set2)
  n1<-length(ind1); n2<-length(ind2)
  w1<-mean(wt[ind1])
  w2<-mean(wt[ind2])
  kernel.mat<-wilcoxon(data$y[ind1],
  data$y[ind2])
  u.hat12<-t(wt[ind1]/w1)%*%kernel.mat%*%
  (wt[ind2]/w2)/(n1*n2)
  ################################################
  #  Calculate variance
  ################################################
  temp1<-(wt[ind1]/w1)%o%rep(1,length(ind2))
  temp2<-rep(1,length(ind1))%o%(wt[ind2]/w2)
  kernel.wt<-temp1*kernel.mat*temp2
  tild.h1<- apply(kernel.wt,1,mean)
  tild.h2<- apply(kernel.wt,2,mean)
  ## Define Cn ##
  Cn <- rep(-Inf, np)
  A<-theta[ind1,]; B<-theta[ind2,]
  #p<-1
  for (p in 1:np){
     C.temp <- outer(A[,p],B[,p],"+")
     CK <- kernel.wt*C.temp
     Cn[p] <- u.hat12*mean(wt[ind1]*A[,p])+
     u.hat12*mean(wt[ind2]*B[,p])-mean(CK)
   }

Cn.3<-t(Cn)%*%jinv%*%t(s)/n.tot  #the third term
```

```
            ### for pasai
            psi1=(-u.hat12*as.vector(t(wt[ind1]/w1)-1)+
            as.vector(apply(kernel.wt, 1,mean)-u.hat12))/n1+
            Cn.3[ind1]
            psi2=(-u.hat12*as.vector(t(wt[ind2]/w2)-1)+
            as.vector(apply(kernel.wt, 2,mean)-u.hat12))/n2+
            Cn.3[ind2]
            #var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2+2*n1*n2*
            mean(outer(psi1, psi2, "*"))
                    var.12<-n1*sd(psi1)^2+n2*sd( psi2)^2
            SET1<-c(SET1,set1)
            SET2<-c(SET2,set2)
            U<-c(U,u.hat12)
            SD<-c(SD,sqrt(var.u[j-1, j-1]))
            P.Val<-c(P.Val, 2*pnorm(-abs(u.hat12-0.5)/
            sqrt(var.12)))
          }
        }
        Pairwise<-cbind(SET1, SET2, U, SD, P.Val)
        colnames(Pairwise)<-c("SET1", "SET2", "U-stat", "SD",
         "P.Val")
        out=list(U.hat=u.hat,Wald.Stat=Stat.Wald,P.val=p.val,
        Pairwise.Comparisons=Pairwise )
        return(out)
 }
wilcoxon_ker <- function(m1,m0){
                    res <- ifelse(m1>m0,1,
                                  ifelse(m1==m0,0.5,0))
                    return(res)
```

```
          }
##################################################
##################################################
##########  GPS based adjustment regression ##########
##################################################
##################################################

RegMethods <- function(data,e,link, method)
{
if(method=="GPSreg")
{
if(link=="probit")
{
         ind1 <- length(unique(g));ind2 <- ind1+1;ind3 <- ind2+1;
         M4 <- clm(factor(y) ~ factor(g)+ e, link="probit",
         data=data)
         v3 <- vcov(M4)[ind1:ind3,ind1:ind3]
         ############################
         ###### Overall Test ######
         ############################

         coef <- summary(M4)$coeff[ind1:ind3,1]
         WT <- wald.test(Sigma=vcov(M4),b=coef(M4),
         Terms=ind1:ind3)
         p.WT <- WT$result[[1]][3]

         ##### Superiority Score Estimates
      sup01 <- pnorm(summary(M4)$coeff[ind1,1]/sqrt(2))
         sup02 <- pnorm(summary(M4)$coeff[ind2,1]/sqrt(2))
```

```
sup03 <- pnorm(summary(M4)$coeff[ind3,1]/sqrt(2))
sup12 <- pnorm((summary(M4)$coeff[ind2,1]-summary(M4)$
coeff[ind1,1])/sqrt(2))
sup23 <- pnorm((summary(M4)$coeff[ind3,1]-summary(M4)$
coeff[ind2,1])/sqrt(2))
sup13 <- pnorm((summary(M4)$coeff[ind3,1]-summary(M4)$
coeff[ind1,1])/sqrt(2))

##### Estimated Variance of superiority scores
Var_sup01 <- summary(M4)$coeff[ind1,2]^(2)*0.5*(
dnorm(summary(M4)$coeff[4,1]/sqrt(2)))^2
Var_sup02 <- summary(M4)$coeff[ind2,2]^(2)*0.5*(
dnorm(summary(M4)$coeff[5,1]/sqrt(2)))^2
Var_sup03 <- summary(M4)$coeff[ind3,2]^(2)*0.5*(
dnorm(summary(M4)$coeff[6,1]/sqrt(2)))^2
Var_sup12 <- 0.5*(dnorm((summary(M4)$coeff[ind2,1]-
summary(M4)$coeff[ind1,1])/sqrt(2)))^2*t(c(-1,1))%*%
v3[c(1,2),c(1,2)]%*%c(-1,1)
Var_sup23 <- 0.5*(dnorm((summary(M4)$coeff[ind3,1]-
summary(M4)$coeff[ind2,1])/sqrt(2)))^2*t(c(-1,1))%*%
v3[c(2,3),c(2,3)]%*%c(-1,1)
Var_sup13 <- 0.5*(dnorm((summary(M4)$coeff[ind3,1]-
summary(M4)$coeff[ind1,1])/sqrt(2)))^2*t(c(-1,1))%*%
v3[c(1,3),c(1,3)]%*%c(-1,1)

##### p-values
p_sup01 <- 2*pnorm(abs((sup01-0.5)/sqrt(Var_sup01)),
lower.tail=FALSE)
p_sup02 <- 2*pnorm(abs((sup02-0.5)/sqrt(Var_sup02)),
```

```
        lower.tail=FALSE)
        p_sup03 <- 2*pnorm(abs((sup03-0.5)/sqrt(Var_sup03)),
        lower.tail=FALSE)
        p_sup12 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup12)),
        lower.tail=FALSE)
        p_sup23 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup23)),
        lower.tail=FALSE)
        p_sup13 <- 2*pnorm(abs((sup13-0.5)/sqrt(Var_sup13)),
        lower.tail=FALSE)
}
if(link=="logit")
{
        ind1 <- length(unique(g));ind2 <- ind1+1;ind3 <- ind2+1;
        M4 <- clm(factor(y) ~ factor(g)+ e, link="logit", data=data)
        v3 <- vcov(M4)[ind1:ind3,ind1:ind3]

        ############################
        ###### Overall Test ######
        ############################

        coef <- summary(M4)$coeff[ind1:ind3,1]
        WT <- wald.test(Sigma=vcov(M4),b=coef(M4), Terms=ind1:ind3)
        p.WT <- WT$result[[1]][3]

        ##### Superiority Score Estimates
      sup01 <- exp(summary(M4)$coeff[ind1,1]/sqrt(2))/(1+
      exp(summary(M4)$coeff[ind1,1]/sqrt(2)))
        sup02 <- exp(summary(M4)$coeff[ind2,1]/sqrt(2))/(1+
        exp(summary(M4)$coeff[ind2,1]/sqrt(2)))
```

```
sup03 <- exp(summary(M4)$coeff[ind3,1]/sqrt(2))/(1+
exp(summary(M4)$coeff[ind3,1]/sqrt(2)))
sup12 <- exp((summary(M4)$coeff[ind2,1]-summary(M4)$
coeff[ind1,1])/sqrt(2))/
(1+exp((summary(M4)$coeff[ind2,1]-
summary(M4)$coeff[ind1,1])/sqrt(2)))
sup23 <- exp((summary(M4)$coeff[ind3,1]-summary(M4)$
coeff[ind2,1])/sqrt(2))/(1+exp((summary(M4)$
coeff[ind3,1]-summary(M4)$coeff[ind2,1])/sqrt(2)))
sup13 <- exp((summary(M4)$coeff[ind3,1]-summary(M4)$
coeff[ind1,1])/sqrt(2))/(1+exp((summary(M4)$
coeff[ind3,1]-summary(M4)$coeff[ind1,1])/sqrt(2)))

##### Estimated Variance of superiority scores
Var_sup01 <- summary(M4)$coeff[ind1,2]^(2)*
exp(summary(M4)$coeff[ind1,1])/(2*(1+exp(
summary(M4)$coeff[ind1,1]))^2)
Var_sup02 <- summary(M4)$coeff[ind2,2]^(2)*
exp(summary(M4)$coeff[ind2,1])/(2*(1+exp(
summary(M4)$coeff[ind2,1]))^2)
Var_sup03 <- summary(M4)$coeff[ind3,2]^(2)*
exp(summary(M4)$        coeff[ind3,1])/(2*(1+exp(
summary(M4)$coeff[ind3,1]))^2)
Var_sup12 <- exp(summary(M4)$coeff[ind2,1]-
summary(M4)$coeff[ind1,1])/(2*(1+exp(summary(M4)$
coeff[ind2,1]-summary(M4)$coeff[ind1,1]))^4)*
t(c(-1,1))%*%v3[c(1,2),c(1,2)]%*%c(-1,1)
Var_sup23 <- exp(summary(M4)$coeff[ind3,1]-
summary(M4)$coeff[ind2,1])/(2*(1+exp(summary(M4)$
```

```
        coeff[ind3,1]-summary(M4)$coeff[ind2,1]))^4)*
        t(c(-1,1))%*%v3[c(2,3),c(2,3)]%*%c(-1,1)
        Var_sup13 <- exp(summary(M4)$coeff[ind3,1]-
        summary(M4)$coeff[ind1,1])/(2*(1+exp(summary(M4)$
        coeff[ind3,1]-summary(M4)$coeff[ind1,1]))^4)*
        t(c(-1,1))%*%v3[c(1,3),c(1,3)]%*%c(-1,1)

        ##### p-values
        p_sup01 <- 2*pnorm(abs((sup01-0.5)/sqrt(Var_sup01)),
        lower.tail=FALSE)
        p_sup02 <- 2*pnorm(abs((sup02-0.5)/sqrt(Var_sup02)),
        lower.tail=FALSE)
        p_sup03 <- 2*pnorm(abs((sup03-0.5)/sqrt(Var_sup03)),
        lower.tail=FALSE)
        p_sup12 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup12)),
        lower.tail=FALSE)
        p_sup23 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup23)),
        lower.tail=FALSE)
        p_sup13 <- 2*pnorm(abs((sup13-0.5)/sqrt(Var_sup13)),
        lower.tail=FALSE)
}
}
if(method=="COVreg")
{
if(link=="probit")
{
        ind1 <- length(unique(g));i
        nd2 <- ind1+1;ind3 <- ind2+1;
        M4 <- clm(factor(y) ~ factor(g) + z.2 + z.3 + z.4 +
```

```
    z.5+ z.6 + z.7 + z.8+ z.9 + z.10 +z.11 + z.12 + z.13,
    link="probit",data=data)
    v3 <- vcov(M4)[ind1:ind3,ind1:ind3]

    ############################
    ###### Overall Test ######
    ############################

    coef <- summary(M4)$coeff[ind1:ind3,1]
    WT <- wald.test(Sigma=vcov(M4),b=coef(M4),
    Terms=ind1:ind3)
    p.WT <- WT$result[[1]][3]

    ##### Superiority Score Estimates
sup01 <- pnorm(summary(M4)$coeff[ind1,1]/sqrt(2))
    sup02 <- pnorm(summary(M4)$coeff[ind2,1]/sqrt(2))
    sup03 <- pnorm(summary(M4)$coeff[ind3,1]/sqrt(2))
    sup12 <- pnorm((summary(M4)$coeff[ind2,1]-summary(M4)$
    coeff[ind1,1])/sqrt(2))
    sup23 <- pnorm((summary(M4)$coeff[ind3,1]-summary(M4)$
    coeff[ind2,1])/sqrt(2))
    sup13 <- pnorm((summary(M4)$coeff[ind3,1]-summary(M4)$
    coeff[ind1,1])/sqrt(2))

    ##### Estimated Variance of superiority scores
    Var_sup01 <- summary(M4)$coeff[ind1,2]^(2)*0.5*(dnorm(
    summary(M4)$coeff[4,1]/sqrt(2)))^2
    Var_sup02 <- summary(M4)$coeff[ind2,2]^(2)*0.5*(dnorm(
    summary(M4)$coeff[5,1]/sqrt(2)))^2
```

```
        Var_sup03 <- summary(M4)$coeff[ind3,2]^(2)*0.5*(dnorm(
        summary(M4)$coeff[6,1]/sqrt(2)))^2
        Var_sup12 <- 0.5*(dnorm((summary(M4)$coeff[ind2,1]-
        summary(M4)$coeff[ind1,1])/sqrt(2)))^2*t(c(-1,1))%*%
        v3[c(1,2),c(1,2)]%*%c(-1,1)
        Var_sup23 <- 0.5*(dnorm((summary(M4)$coeff[ind3,1]-
        summary(M4)$coeff[ind2,1])/sqrt(2)))^2*t(c(-1,1))%*%
        v3[c(2,3),c(2,3)]%*%c(-1,1)
        Var_sup13 <- 0.5*(dnorm((summary(M4)$coeff[ind3,1]-
        summary(M4)$coeff[ind1,1])/sqrt(2)))^2*t(c(-1,1))%*%
        v3[c(1,3),c(1,3)]%*%c(-1,1)

        ##### p-values
        p_sup01 <- 2*pnorm(abs((sup01-0.5)/sqrt(Var_sup01)),
        lower.tail=FALSE)
        p_sup02 <- 2*pnorm(abs((sup02-0.5)/sqrt(Var_sup02)),
        lower.tail=FALSE)
        p_sup03 <- 2*pnorm(abs((sup03-0.5)/sqrt(Var_sup03)),
        lower.tail=FALSE)
        p_sup12 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup12)),
        lower.tail=FALSE)
        p_sup23 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup23)),
        lower.tail=FALSE)
        p_sup13 <- 2*pnorm(abs((sup13-0.5)/sqrt(Var_sup13)),
        lower.tail=FALSE)
}
if(link=="logit")
{
        ind1 <- length(unique(g));ind2 <- ind1+1;
```

```
    ind3 <- ind2+1;
    M4 <- clm(factor(y) ~ factor(g)+ z.2 + z.3 +
    z.4 + z.5+ z.6 + z.7 + z.8+ z.9 + z.10 +
    z.11 + z.12 + z.13, link="logit", data=data)
    v3 <- vcov(M4)[ind1:ind3,ind1:ind3]

    ############################
    ###### Overall Test ######
    ############################

    coef <- summary(M4)$coeff[ind1:ind3,1]
    WT <- wald.test(Sigma=vcov(M4),b=coef(M4),
    Terms=ind1:ind3)
    p.WT <- WT$result[[1]][3]

    ##### Superiority Score Estimates
sup01 <- exp(summary(M4)$coeff[ind1,1]/sqrt(2))/
(1+exp(summary(M4)$coeff[ind1,1]/sqrt(2)))
    sup02 <- exp(summary(M4)$coeff[ind2,1]/sqrt(2))/
    (1+exp(summary(M4)$coeff[ind2,1]/sqrt(2)))
    sup03 <- exp(summary(M4)$coeff[ind3,1]/sqrt(2))/
    (1+exp(summary(M4)$coeff[ind3,1]/sqrt(2)))
    sup12 <- exp((summary(M4)$coeff[ind2,1]-summary(M4)$
    coeff[ind1,1])/sqrt(2))/(1+exp((summary(M4)$
    coeff[ind2,1]-summary(M4)$coeff[ind1,1])/sqrt(2)))
    sup23 <- exp((summary(M4)$coeff[ind3,1]-summary(M4)$
    coeff[ind2,1])/sqrt(2))/(1+exp((summary(M4)$
    coeff[ind3,1]-summary(M4)$coeff[ind2,1])/sqrt(2)))
    sup13 <- exp((summary(M4)$coeff[ind3,1]-summary(M4)$
```

```
coeff[ind1,1])/sqrt(2))/(1+exp((summary(M4)$
coeff[ind3,1]-summary(M4)$coeff[ind1,1])/sqrt(2)))

##### Estimated Variance of superiority scores
Var_sup01 <- summary(M4)$coeff[ind1,2]^(2)*exp(
summary(M4)$coeff[ind1,1])/(2*(1+exp(summary(M4)$
coeff[ind1,1]))^2)
Var_sup02 <- summary(M4)$coeff[ind2,2]^(2)*exp(
summary(M4)$coeff[ind2,1])/(2*(1+exp(summary(M4)$
coeff[ind2,1]))^2)
Var_sup03 <- summary(M4)$coeff[ind3,2]^(2)*exp(
summary(M4)$coeff[ind3,1])/(2*(1+exp(summary(M4)$
coeff[ind3,1]))^2)
Var_sup12 <- exp(summary(M4)$coeff[ind2,1]-summary(M4)$
coeff[ind1,1])/(2*(1+exp(summary(M4)$coeff[ind2,1]-
summary(M4)$coeff[ind1,1]))^4)*t(c(-1,1))%*%v3[c(1,2),
c(1,2)]%*%c(-1,1)
Var_sup23 <- exp(summary(M4)$coeff[ind3,1]-summary(M4)$
coeff[ind2,1])/(2*(1+exp(summary(M4)$coeff[ind3,1]-
summary(M4)$coeff[ind2,1]))^4)*t(c(-1,1))%*%v3[c(2,3),
c(2,3)]%*%c(-1,1)
Var_sup13 <- exp(summary(M4)$coeff[ind3,1]-summary(M4)$
coeff[ind1,1])/(2*(1+exp(summary(M4)$coeff[ind3,1]-
summary(M4)$coeff[ind1,1]))^4)*t(c(-1,1))%*%v3[c(1,3),
c(1,3)]%*%c(-1,1)

##### p-values
p_sup01 <- 2*pnorm(abs((sup01-0.5)/sqrt(Var_sup01)),
lower.tail=FALSE)
```

```
        p_sup02 <- 2*pnorm(abs((sup02-0.5)/sqrt(Var_sup02)),
        lower.tail=FALSE)
        p_sup03 <- 2*pnorm(abs((sup03-0.5)/sqrt(Var_sup03)),
        lower.tail=FALSE)
        p_sup12 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup12)),
        lower.tail=FALSE)
        p_sup23 <- 2*pnorm(abs((sup12-0.5)/sqrt(Var_sup23)),
        lower.tail=FALSE)
        p_sup13 <- 2*pnorm(abs((sup13-0.5)/sqrt(Var_sup13)),
        lower.tail=FALSE)
}
}
out=list(coef,v3,p.WT,sup01,sup02,sup03,sup12,sup23,sup13,
        Var_sup01,Var_sup02,Var_sup03,Var_sup12,Var_sup23,
        Var_sup13,p_sup01,p_sup02,p_sup03,p_sup12,p_sup23,p_sup13)
        names(out)=c("coef","coef_var","p.WT","sup01","sup02",
"sup03","sup12","sup23","sup13","Var_sup01","Var_sup02",
"Var_sup03","Var_sup12","Var_sup23","Var_sup13","p_sup01",
"p_sup02","p_sup03","p_sup12","p_sup23","p_sup13")
        return(out)
}
########################################
########## Stratified U stat ##########
########################################

wilcoxon_var <- function(y1,y0){
        m1 <- length(y1); m0 <- length(y0);
        n <- m1+m0
        mat <- mat.or.vec(m1,m0)
```

```
        for(i in 1:m1)
        {
                for(j in 1:m0)
                {
                mat[i,j] <- ifelse(y1[i]>y0[j],1,
                ifelse(y1[i]==y0[j],0.5,0))
                }
        }
theta.hat <- mean(mat)
h10 <- rowMeans(mat)-theta.hat
h01 <- colMeans(mat)-theta.hat

A1 <- (1/(m1-1))*sum((h10-theta.hat)^2)
B1 <- (1/(m0-1))*sum((h01-theta.hat)^2)
var <- A1/m1+B1/m0
return(var)
}
stratU <- function(data)
{
  sup01 <- ifelse(all(c(1,2) %in% unique(data$g)),
 mean(outer(data$y[data$g==2],
 data$y[data$g==1],">"))+0.5*mean(outer(data$y[data$g==2],
 data$y[data$g==1],"==")),NA)
 sup02 <- ifelse(all(c(1,3) %in% unique(data$g)),
 mean(outer(data$y[data$g==3],
 data$y[data$g==1],">"))+0.5*mean(outer(data$y[data$g==3],
 data$y[data$g==1],"==")),NA)
 sup03 <- ifelse(all(c(1,4) %in% unique(data$g)),
 mean(outer(data$y[data$g==4],
```

```
 data$y[data$g==1],">"))+0.5*mean(outer(data$y[data$g==4],
 data$y[data$g==1],"==")),NA)
 sup12 <- ifelse(all(c(2,3) %in% unique(data$g)),
 mean(outer(data$y[data$g==3],
 data$y[data$g==2],">"))+0.5*mean(outer(data$y[data$g==3],
 data$y[data$g==2],"==")),NA)
 sup23 <- ifelse(all(c(3,4) %in% unique(data$g)),
 mean(outer(data$y[data$g==4],
 data$y[data$g==3],">"))+0.5*mean(outer(data$y[data$g==4],
 data$y[data$g==3],"==")),NA)
 sup13 <- ifelse(all(c(2,4) %in% unique(data$g)),
 mean(outer(data$y[data$g==4],
 data$y[data$g==2],">"))+0.5*mean(outer(data$y[data$g==4],
 data$y[data$g==2],"==")),NA)

 varS_sup01 <- wilcoxon_var(data$y[data$g==1],data$y[data$g==2])
 varS_sup02 <- wilcoxon_var(data$y[data$g==1],data$y[data$g==3])
 varS_sup03 <- wilcoxon_var(data$y[data$g==1],data$y[data$g==4])
 varS_sup12 <- wilcoxon_var(data$y[data$g==2],data$y[data$g==3])
 varS_sup23 <- wilcoxon_var(data$y[data$g==3],data$y[data$g==4])
 varS_sup13 <- wilcoxon_var(data$y[data$g==2],data$y[data$g==4])

out= list(sup01,sup02,sup03,sup12,sup23,sup13,varS_sup01,
        varS_sup02, varS_sup03, varS_sup12, varS_sup23, varS_sup13)
names(out) <- c("sup01","sup02","sup03","sup12","sup23","sup13",
"Var_sup13")
return(out)
}
```

A.8 R code for Generalized Spatiotemporal Additive Model Implemented in R and Its Application to Assessing Overuse of Antibiotics Drugs for Upper Respiratory Tract Infections in Kentucky

Important portions of R codes to analyze the generalized spatiotemporal additive model in Chapter 4 are provided below. **R code**:

```
 rm(list=ls())
#setwd("F:/Project 3")
library(mgcv)
library(visreg)
library(plyr)
library(ggplot2)
library(dplyr)
library(lme4)

################################################
######### URIdata is the final data ########
################################################

URIdata <- readRDS("URI_allzip_final.rds", refhook=NULL)
str(URIdata)

################################################
######### Trend components ####################
################################################

URIdata$m <- NA

URIdata$m[which(URIdata$Year=="2014" &
```

```
URIdata$Service_Month=="January")] <- 1
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="February")] <- 2
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="March")] <- 3
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="April")] <- 4
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="May")] <- 5
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="June")] <- 6
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="July")] <- 7
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="August")] <- 8
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="September")] <- 9
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="October")] <- 10
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="November")] <- 11
URIdata$m[which(URIdata$Year=="2014" &
URIdata$Service_Month=="December")] <- 12

URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="January")] <- 13
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="February")] <- 14
URIdata$m[which(URIdata$Year=="2015" &
```

```
URIdata$Service_Month=="March")] <- 15
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="April")] <- 16
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="May")] <- 17
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="June")] <- 18
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="July")] <- 19
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="August")] <- 20
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="September")] <- 21
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="October")] <- 22
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="November")] <- 23
URIdata$m[which(URIdata$Year=="2015" &
URIdata$Service_Month=="December")] <- 24

URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="January")] <- 25
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="February")] <- 26
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="March")] <- 27
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="April")] <- 28
URIdata$m[which(URIdata$Year=="2016" &
```

```
URIdata$Service_Month=="May")] <- 29
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="June")] <- 30
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="July")] <- 31
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="August")] <- 32
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="September")] <- 33
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="October")] <- 34
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="November")] <- 35
URIdata$m[which(URIdata$Year=="2016" &
URIdata$Service_Month=="December")] <- 36

URIdata$cos2pi <- cos(2*pi*URIdata$m/12)
URIdata$cos4pi <- cos(4*pi*URIdata$m/12)
URIdata$cos6pi <- cos(6*pi*URIdata$m/12)
URIdata$cos8pi <- cos(8*pi*URIdata$m/12)
URIdata$cos10pi <- cos(10*pi*URIdata$m/12)
URIdata$cos12pi <- cos(12*pi*URIdata$m/12)

URIdata$sin2pi <- sin(2*pi*URIdata$m/12)
URIdata$sin4pi <- sin(4*pi*URIdata$m/12)
URIdata$sin6pi <- sin(6*pi*URIdata$m/12)
URIdata$sin8pi <- sin(8*pi*URIdata$m/12)
URIdata$sin10pi <- sin(10*pi*URIdata$m/12)
URIdata$sin12pi <- sin(12*pi*URIdata$m/12)
```

```
URIdata$year_cont <- URIdata$m/12
prac <- na.omit(URIdata)

#######################################
########### bam
#######################################

tm <- proc.time()
URI_bam <- bam(Antibacterial~s(Age, bs="cr")+Sex+Race+
        Region_Code+s(Pediatrician_per10000, bs="cr")+
        s(Unemploy_rate, bs="cr")+s(Perc_poverty, bs="cr") +
        year_cont+ cos2pi+cos4pi+cos6pi+cos8pi+cos10pi+
        sin2pi+sin6pi+sin8pi+sin10pi+
        s(zip_lat, zip_long, bs="tp")+Provider_NPI_ID,
        #sp=c(URI_prac$zip_lat, URI_prac$zip_long),
        data = URIdata, family = binomial(link = "logit"))
proc.time()-tm
## 141 secs (n=10000), 16.33 hrs (n=455269)
summary(URI_bam)
s <- summary(URI_bam)
```

CURRICULUM VITA

NAME: Soutik Ghosal

ADDRESS: Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, KY 40292

EDUCATION: Bachelor of Science in Statistics,
Presidency College, India, 2012
Master of Science in Statistics,
Presidency University, India, 2014

PUBLICATIONS: Davis, D. W., Feygin, Y.,. . . , Ghosal, S.,. . . , McKinley, S.(2018). Longitudinal trends in ADHD diagnosis and stimulant use in preschool children on Medicaid.
Journal of Pediatrics (Submitted).

Ghosal, S., Lau, T. S., Gaskins, J. & Kong, M. (2018). Hierarchical Mixed Effect Hurdle Model for Time and Hierarchical mixed effect hurdle model for time and spatially correlated count data and its application to identifying factors impacting health professional shortages.
Journal of the Royal Statistical Society: Series C

(Applied Statistics) (Submitted).

Nessle, C. N., Ghosal, S., Mathews, C., Taylor, D., Myers, J., Raj, A.& Panigrahi, A. (2018).
Weak correlation of bleeding scores to platelet electron microscopy: A retrospective chart review of 109 pediatric patients with δ-storage pool disorder.
Pediatric Blood & Cancer(Submitted).

Ghosal, S., Trivedi, J., Chen, J., Rogers, M. P., Cheng, A., Slaughter, M. S.,Kong, M. & Huang, J. (2017).
Regional Cerebral Oxygen Saturation Level Predicts 30-Day Mortality Rate After Left Ventricular Assist Device Surgery.
Journal of Cardiothoracic and Vascular Anesthesia.

Ghosal, S., Kong, M. (in process)
Assess treatment effects for multiple groups when outcome is ordinal and confounding variables exist.

Ghosal, S., Myers, J., Smith, M. J., Kong, M.
Generalized Spatiotemporal Additive Model and Its Application to Assessing Overuse of Antibiotics Drugs for Upper Respiratory Tract Infections in Kentucky (in process).

Ghosal, S., Trivedi, J., Barlowe, D., ... Slaughter, M. S., Kong, M. & Huang, J.(in process)
Preoperative Coagulation Parameters Predict 30-Day

Mortality after Cardiac Surgery- A Retrospective Study.

PRESENTATIONS: Graduate Student Regional Research Conference. March 3, 2018.
Generalized Spatiotemporal Additive Model and Its Application to Assessing Overuse of Antibiotics Drugs for Upper Respiratory Tract Infections in Kentucky.

Kentucky ASA Chapter Meeting. March 2, 2018.
Assess Treatment Effects for Multiple Groups for Ordinal Outcome when Confounding Exists.

Department of Bioinformatics & Biostatistics Seminar Series. January 26, 2018.
Spatiotemporal mixed effect modeling and its applications.

Joint Statistical Meetings (JSM). August 30, 2017.
Hierarchical Mixed Effect Hurdle Model for Time and Spatially Correlated Count Data and its Bayesian Analysis.

Kentucky ASA Chapter Meeting. April 21, 2017.
Hierarchical Mixed Effect Hurdle Model for Time and Spatially Correlated Count Data and its Bayesian Analysis.

HONORS AND AWARDS NSF funded Harshbarger Travel award for the SRCOS Summer Research Conference, Jekyll Island, GA, 2017

NSF funded Harshbarger Travel award for the SRCOS Summer Research Conference, Virginia Beach, VA, 2018

School of Public Health Travel Scholarship (University of Louisville) for ENAR January 2018